DOCUMENT RESUME

ED 193 950 FL 011 928

AUTHOR Goodrich, Robert L.; And Others
TITLE Planning Factors for Studies of Bilingual Instructional Features. Planning Paper 3. Bilingual Instructional Features Planning Study.
INSTITUTION Abt Associates, Inc., Cambridge, Mass.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C. National Council on Educational Research.
REPORT NO AAI-80-4
PUB DATE 1 Jan 80
CONTRACT 400-79-0071
NOTE 114p.; For related documents, see FL 011 926-930.

EDRS PRICE MF01/PC05 Plus Postage.
DESCRIPTORS *Bilingual Education; *Educational Planning; Educational Research; Elementary Secondary Education; *Instructional Development; *Research Design; Research Methodology; *Sampling; Testing
IDENTIFIERS National Institute of Education

ABSTRACT

This report is one of five submitted as products of a "Study Designed to Assist Planning of Research on Significant Instructional Features in Bilingual Education Programs." The reports are intended to assist the National Institute of Education (NIE) in its plans for a major new research study in bilingual education. The present report is based on a review of studies of educational instructional features in both monolingual and bilingual contexts, and on conversations with a large number of researchers and critics. The intent of the report is to summarize the state of the art of bilingual educational features research as a base for designs to be developed by NIE. The design according to which sample programs were chosen for the study is set forth. Issues in the analytic study of population groups targeted for bilingual education are discussed. A chapter is devoted to measurement issues. The volume concludes with a discussion of the management, organization, and phasing of the study. A bibliography is appended. (Author/JB)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ABT ASSOCIATES INC.
55 WHEELER STREET, CAMBRIDGE, MASSACHUSETTS 02138
TELEPHONE • AREA 617-492-7100

ED193950

AAI Report #80-4

BILINGUAL INSTRUCTIONAL FEATURES PLANNING STUDY

Planning Paper 3

Planning Factors for Studies of Bilingual Instructional Features

by

Robert L. Goodrich

with
contributions by

Gaea Leinhardt
Edward Cervenka
Jose Llanes
Robert Carrasco

January 1, 1980

Contract No. NIE-400-79-0071
Study Designed To Assist Planning of Research on Significant Instructional Features in Bilingual Education Programs: Task 3 Report

Robert L. Goodrich, Project Director
Sarah Nieves-Squires, Deputy Project Director

Mae Chu-Chang, Project Officer

Prepared for:
The National Institute of Education
1200 19th St. N.W.
Washington, D.C. 20208

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

FL011928

Contract Manager	Quality Control Reviewer	Management Reviewer

SEP 2 1980

FOREWORD

This report is one of five submitted by Abt Associates Incorporated (AAI) to the National Institute of Education, as products of a "Study Designed to Assist Planning of Research on Significant Instructional Features in Bilingual Education Programs", Contract No. NIE-400-79-0071. The reports are intended to assist NIE in its plans for a major new research study in bilingual education. The information provided will be combined with that from other sources by NIE in its construction of a research plan, to be incorporated in one or more requests for proposals (RFP's) to implement and conduct the major study.

The Instructional Features Study was formulated by the Division of Education Part C Coordinating Committee as one of several studies that implement research mandates in the language of ESEA Title VII, Part C. A description of the study (denoted "B-1") is provided in the U.S.H.E.W. Research Plan for Bilingual Education (July, 1979). This planning assistance study was one component of Phase I of the Three phase HEW plan.

These reports were prepared as products of Tasks 1-5 of the planning assistance contract. The titles of the reports, and summaries of their contents, are:

1. Working Definitions of Terms for the Bilingual Instructional Features Study, by Sarah Nieves-Squires and Robert L. Goodrich.

   This is a discussion of working definitions of terms for use in the features study. The terms discussed are "bilingual education", "consequences for children", "instructional features", "significant", and "model". Alternative definitions and the implications of each for design are presented. The definitions selected by NIE are intended to guide the research to be conducted.

2. A Bibliography of Significant Instructional Features in Bilingual Education Programs, by Sarah Nieves-Squires, et al.

   This is an annotated bibliography of papers, articles, pamphlets and books that deal with instructional features of bilingual education. The materials are organized by a classification system of features based on a content analysis of the sources surveyed. The report demonstrates that, while many instructional features are discussed, there is little or no empirically based research on their specific consequences for children.

3. Planning Factors for Studies of Bilingual Instructional Features, by Robert L. Goodrich.

   This report is based on a review of studies of educational instructional features in both monolingual and bilingual contexts, and on conversations with a large number of researchers and critics. The intent of the report is to summarize the state of the art of bilingual educational features research as a base for designs to be developed by NIE.

4. Tentative Alternative Designs for a Study of Significant Instructional Features in Bilingual Education, by Robert L. Goodrich.

   This report presents alternative study designs and plans for implementation of the instructional features study. It is based on the knowledge base assembled in the three preceding reports. The designs presented are not to be implemented directly by NIE in the RFP, but used. Rather, they are simply one source of information available to the NIE planners, to be factored into the overall design process.

5. Feasibility and Credibility of Bilingual Instructional Features Study Plans: Field Verification, by Sarah Nieves-Squires, Robert L. Goodrich, and Cristina Bodinger-de Uriarte.

   This report summarizes the results of 123 open interviews conducted with bilingual practitioners and administrators in five sites: Los Angeles, Miami, New York, Oakland, Rough Rock, AZ. The questions asked were designed to elicit responses about the working definitions of terms (Paper 1) and the designs considered (Papers 3, 4). The intent was to test the credibility and acceptability to consumers of alternative study approaches.

The competition between paradigms is not the sort of battle that can be resolved by proofs.....The transfer of allegiance from paradigm to paradigm is a conversion experience that cannot be forced.

-Thomas S. Kuhn

Table of Contents

		Page
1	INTRODUCTION	1
2	SAMPLE DESIGN	4
2.1	Definition of Sample Frame	4
2.2	Selection of Stratifiers	13
2.3	Site Selection	15
2.4	Sample Sizes and Statistical Power	16
2.5	Sampling Method	18
2.6	Adjustability of Scale	19
3	ANALYTIC STUDIES	20
3.1	Merging Quantitative and Qualitative Paradigms	21
3.2	Diversity of Cultural/Linguistic Groups	36
3.3	Linguistic Proficiency Issues	43
3.4	Analytic Subsamples of Children	44
3.5	The Role of Comparison Groups.	45
3.6	Contextual Effects	47
3.7	Teacher/Aide Qualifications	48
3.8	Definition of Treatment.	50
3.9	Generalizability	51
3.10	The Time Frame and Measurement of Change	52
3.11	Longitudinal Studies	54
4	MEASUREMENT ISSUES	57
4.1	Classroom Process.	58
4.1.1	Alternative Measurement Techniques	59
4.2	Measures of Language Proficiency and Dominance	79
5	MANAGEMENT, ORGANIZATION AND PHASING	88
5.1	Organization of Study Components	88
5.2	Management	89
5.3	Phasing.	95
Bibliography.		99

Chapter 1
INTRODUCTION

This report is the third in a sequence prepared for the National Institute of Education to assist it in planning new research studies of instructional features of bilingual education (BE). The first two reports present discussions of definitions of terms in bilingual education, and accumulated knowledge about instructional features in bilingual education, respectively. The purpose of this report is to "identify what is already known about factors to be considered in designing a new study and lessons from the previous efforts regarding trade-offs in site selections, sample sizes, ways of merging qualitative and quantitative information, time and costs of the study, management and consortium approaches, and all other features of design that could influence success in this particular study." The results presented in the first three reports will be used directly in preparing the fourth, where alternative study designs will be set forth. These alternative designs, and the discussions presented in the other planning reports, will be factored by NIE into its own planning process.

The potentially enormous scope of the task to be presented here is obvious from the statement of purpose presented above. We cannot, of course, present all knowledge in the areas identified. Another researcher might make different emphases than our own. The methods and strategies used to prepare the report are therefore of some importance in placing it in proper perspective as the author's opinions, grounded as we discuss below.

Most of the general design considerations discussed below come from the author's experience in interpreting and designing educational and psychological studies over a period of years with Abt Associates Incorporated (AAI). Specific Bilingual Education instructional features

design considerations came from several sources: (1) a reading of selected important research reports and reviews in the fields of bilingual education, instructional dimensions research, sociolinguistics, ethnography and microethnography; (2) separate technical conversations with a large number of researchers in these fields (listed in an appendix); (3) technical conversations with colleagues at the Abt Associates Incorporated, Cambridge offices; and (4) use of extensive literature abstracts from academic and non-academic sources prepared by George Blanco (University of Texas) and Jose Llanes (Eastern Michigan University), with the support of graduate student consultants (to AAI) working under them.

As many analysts (e.g. Tucker, 1979; Troike, 1978) have noted, the state of the art in Bilingual Education research is not well advanced. Only one large scale study has been conducted and that one was seriously flawed. Of smaller scale evaluations, only a few satisfy minimal methodological criteria. Furthermore, the reported results have been mostly negative or neutral. In general, bilingual education research and evaluation have not contributed substantially to promoting bilingual education as a partial solution to the social and educational problems that led to its formulation. This is true even though a systematic reading of the literature leads the analyst to conclude that bilingual education indeed has considerable positive potential if properly defined and implemented. This circumstance leads to a major responsibility, and a considerable technical challenge, in designing new research. It seems essential to avoid the large scale research study that, because of its exclusive reliance on the quantitative paradigm, is likely to miss the positive effects of bilingual education. While it appears that the total scale of the proposed study must be large, in order to provide all the information required, it should consist of a cohesive mixture of large and small scale, quantitative and qualitative, components.

This report is not intended to provide any definite conclusions. Its sole purpose is to identify and discuss factors that should be considered in planning new bilingual education instructional features studies. Later

reports will present a series of alternative designs based on these factors. The purpose of these reports will be to assist NIE in its own planning and not to provide a single definitive plan.

Nor is this report intended to be part of a linear and convergent process that begins with specific research questions and procedes to a design. Rather, it is intended to raise questions as to what should be studied and what specific research questions might be addressed, and to discuss the design considerations that arise from the questions. While the research questions should not be rigidly limited by measurement technology available, the technology does limit the studies that could be feasibly implemented. Therefore the report should include a survey of the existing methodology and opportunities for using it in a large scale study. The report begins with these technical considerations.

This report is organized by domains of research factors to be considered. These include sample design; analytic issues; measurement issues; and organization, management, and phasing of studies; these topics are covered in Chapters 2 through 5, respectively.

Chapter 2
SAMPLE DESIGN

The bilingual instuctional features study will be based on a sample (or several samples) of bilingual education projects for study. The characteristics of the sample plans are crucial for the credibility and generalizability of study results. This section discusses issues that must be considered in formulating a plan. Specific issues considered are as follows:

- o Definition of Sample Frame
- o Selection of Stratifiers
- o Site Selection
- o Sample Size
- o Sampling Method (purposive, randomized, controlled selection)
- o Adjustibility of Scale

Each of these issues is discussed in the following pages.

2.1 Definition of Sample Frame

One factor that emerges from the discussion of analytic design factors in Chapter 3 bilingual education is that instructional features studies should comprise a multidisciplinary collection of interlinking studies and substudies. These should fit into a mosaic pattern that makes available the overall knowledge needed of the instructional process. These can concentrate on children, parents, teachers, projects, schools or any of a number of other units of analysis. Figure 1 provides a simplified diagram of the hierarchical sample frame for bilingual education instructional features studies. It is obvious that an overall multistage sample plan should link all major studies and substudies. Subsamples for special

purpose studies would be selected within this framework. The first decision to be made in defining the overall sample frame is to specify its scope. Decisions about sampling units and detailed plans for samples and subsamples will follow.

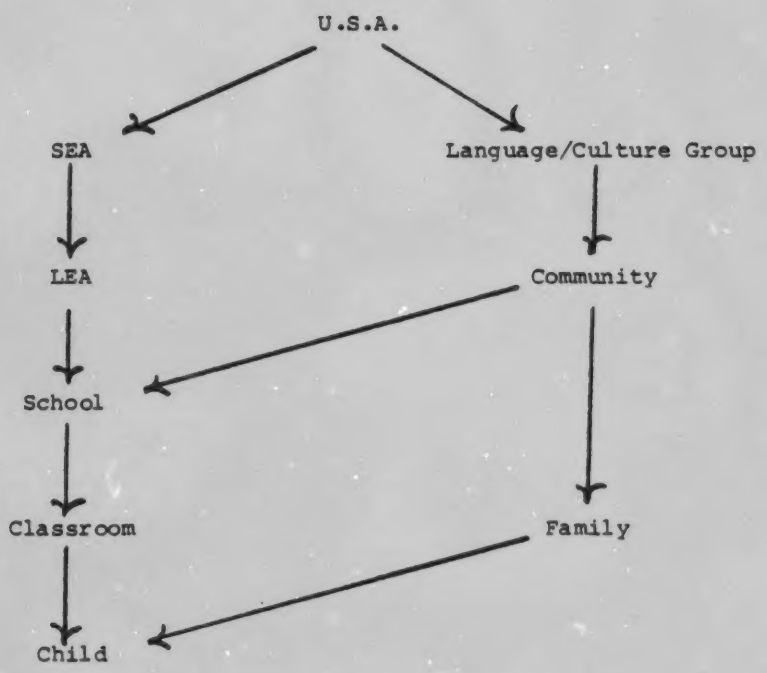


Figure 1

Diagram of Hierarchical Sample Frame

The scope of the sample frame indicates the largest universe of generalizability of study results, i.e. the population to which results can be applied. Of course, focused substudies supporting the main study might concentrate on particular subuniverses and would not be generalizable to the entire sample frame, but they should fit into this frame in a logical way. As will be seen from the discussion below, selection of the sample frame entails an overall approach to studies to be conducted.

In considering the different possibilities, we found three alternatives for defining the overall study universe. These alternatives, in increasing order of generality and scope, are

(1) the universe of Title VII projects,

(2) the universe of projects that satisfy the operational definition of bilingual education developed in working paper 1, and

(3) the universe of schools with analytically significant numbers of linguistic minority children.

We discuss these alternatives in the paragraphs below and indicate the advantages and disadvantages of each.

Alternative Sample Scopes

The three sample scopes to be discussed are based on three alternative definitions of the term "bilingual education" presented in Planning Paper 1 (Working Definition for Studies of Bilingual Education Instructional Features.) These definitions are of progressively larger scope.

(1) Universe of Title VII Projects

This universe contains approximately one-third of the total number of classrooms devoted to bilingual education. It is clearly the most

significant sample from the point of view of the public law, but it would lack significance in relationship to potential instructional features which do not closely follow the mandated qualities of the Title VII programs. While there is much diversity in the sample, and Title VII is itself an experimental demonstration act, the consensus appears to be that the most innovation, and some of the most interesting instructional characteristics, lie outside the universe of Title VII projects.

The knowledge to be produced by NIE studies on bilingual education instructional features should be broadly applicable and more research oriented, and should certainly not compete with the USOE (Evaluation of the Classroom Instructional Component of the ESEA Title VII Bilingual Education Program) study in producing the same impact evaluations.

The theme of NIE studies should be basic research on bilingual education instructional features, conducted by a multidisciplinary examination of natural variations in a study sample of sites and projects. These research objectives are best pursued by examining the entire range of program variations, both inside and outside Title VII. The Title VII universe of projects is, in other words, too constrained to support a research study of the scope that appears desirable.

(2) Universe of Bilingual Education Projects

A more appealing alternative would be to include in the sample frame all projects that satisfy the Second Definition of bilingual education presented in Planning Paper 1 of this planning assistance study. Numerous additional state, private, and locally funded projects would be included and the total scope broadened substantially. A strong research study could be based on this sample frame. This approach is problematic for the following reasons.

First, the sample frame would be very difficult to operationalize. This would involve a large scale front end effort in collecting project particulars from state and local authorities and even then, it would be hard to determine precisely the types of projects involved.

Second, this sample frame would not permit us to compare educational features in bilingual and monolingual settings, unless monolingual comparison schools were added to the sample. The difficulties in locating comparison schools with very many limited English Proficiency (LEP) children are discussed in Chapter 3 below.

Third, a weakness of using the sample frame of bilingual education projects is that it does not allow for the fact that only a part of language education takes place in the school setting. Family, church, and community all influence the language development. In fact there are many programs in community institutions (churches, temples, ethnic associations) that deliver formal or informal instruction for the linguistic minority child after school and on weekends. These should be included in any comprehensive examination of the education of the linguistic minority child.

(3) The Universe of Schools with Linguistic Minority Students

The third alternative sample frame concentrates on the linguistic minority student - i.e. the potential consumer of bilingual education - rather than the program. The Office for Civil Rights biennial survey (discussed below) affords a practical way to implement the most complete sample alternative. In this design we would essentially sample students and progress analytically from the student to his or her own entire support system, including the public educational component. The sample frame may include all schools where meaningful numbers of such children are enrolled, whether or not they have bilingual education projects. The formal sampling unit can be either the school or the LEA. Thus, analyses of instructional

features can be carried across bilingual and monolingual approaches and the distinctive features of bilingual education identified. The major limitation of the OCR Survey is that it does not include private and parochial schools. The availability and nature of the OCR data source is described below.

Potential Use of OCR Survey Data

An outstanding resource in defining an overall sample frame of schools or school districts is provided by the OCR survey conducted (on a large scale) every two years. The 1978-79 survey should be available in time for use in BE instructional features studies, but the information given below is from the 1976-77 publication (DHEW/OCR, ND). We have also discussed the OCR sample plan and accessibility of files with Neil Killilea of Killilea Associates (the current contractor to OCR) and Don Reisler of DBS Corporation (the previous contractor).

The 1976-77 Elementary and Secondary Schools Civil Rights Survey consisted of two survey instruments:

1. School System Summary Report (Form OCR/101).
2. Individual School Campus Report (Form OCR/102).

The OCR/101 form was administered to all districts (i.e., LEA's) in the U.S.A. In addition, a sample of approximately 3,600 districts was selected to receive OCR/102 forms. All elementary and secondary schools in a selected district were required to complete the OCR/102 form. Even though the sample of districts was selected by a weighted random sampling procedure that concentrated on districts of "high OCR interest", the sample was drawn so that state-by-state and national projections could be made.

For each district in the country, the following data relevant to the sample definitions are available.

(1) Numbers of pupils with primary or home language other than English, which these pupils speak exclusively or speak more often than English. These data, it should be remarked, are not always reliable and should be used only as a guide in making selections.

(2) Numbers of the above who are in ESL programs or in any class taught in a language other than English.

Each of these groups is broken down into the following ethnic groups.
American Indian or Alaskan Native
Asian or Pacific Islander
Black, not of Hispanic Origin
White, not of Hispanic Origin
Hispanic

Within the subsample of 3600 districts these data are available at school level.

The 1978-79 survey is being conducted by Killalea Associates, Arlington, Virginia. Data will be available in machine-readable form and could serve directly as a sample frame for selection of districts or schools.

Note that since the OCR school level survey is a weighted probability sample, where each district had a positive probability of selection, results from the survey are generalizable to the public school universe. Moreover, a school level subsample from this file would remain generalizable -- i.e., even though we would deal with a sample of a sample the results can be made equivalent to a single stage weighted sample. This data base appears to be desirable for the sampling scheme proposed.

Selection of the Primary Sampling Unit (PSU)

In a multistage sample of the type envisioned for the instructional features studies, the primary sampling unit (PSU) is the first of a sequence of nested (or parallel) sample units (primary, secondary, tertiary,...),

each with its own well defined sample frame. The determination of the overall scope of the sample frame does not necessarily determine the PSU. For instance, if the scope is to consist of the Title VII universe, then we could still use either the project, the school, or the classroom as PSU. There are several possible choices for PSU, including the school project, LEA, community, or metropolitan area. If the OCR survey data base is used to define the sample frame, then the most convenient choices are the school, district, or metropolitan area.

Planning Papers 1 and 2 emphasize the need to study bilingual education in its community context. Each school in the study sample should be considered to be part of a community, that requires macro-ethnographic and anthropological study. The most efficient approach to such a nested (classroom in school in community) study plan is to select the metropolitan area or county as the PSU and to draw subsamples within it. This would enable study of variations in instructional features across several schools embedded within a single community context. A further advantage is that site offices could be more efficiently organized through centralization. A disadvantage is that some generalizability to the public schools programmatic universe could be lost through site clustering of the type suggested.

Selection of the LEA as PSU can offer similar advantages and disadvantages depending on the size of the LEA, its relation to ethnic communities, etc. Obviously the LEA is a good choice as PSU for sub-studies that concentrate on organization, money flow, implementation, etc.

The selection of school as PSU is less efficient conceptually and more expensive, but affords increased generalizability to the universe of schools. If the OCR files are used to define the data base then the OCR/102 school level subsample must be used. (The OCR/101 survey is district level.) This is a small disadvantage but still one that should be considered.

Subpopulations of Students

No matter what the PSU, the sample should be consumer oriented. The populations of students to which results can be generalized should be clearly identified and should include a sizeable subsample of LEP students in grades K-3. A recent Title VII audit showed that about 70% of the students were LEP, up from about 30% in 1976. It is not known what fraction of children in Bilingual Education Programs outside Title VII are LEP. If the sample of LEA's or schools is identified via the OCR survey file, then it would be easy to select only those projects with analytically adequate numbers of LEP children.

If language proficiency testing (L1, L2) is adopted--and this is highly desirable--then a number of interesting analytic subsamples can be identified. The distribution of (L1, L2) proficiencies in a bilingual classroom is one of its most interesting features, and is one consequence of the operational entry-exit system in use.

Grades to be Included

Most studies have concentrated on grades K-6 and sometimes only on grades K-3. Consequences for children do not cease at grade 6, however, and even extend into post-secondary years. Haro (1977) discusses a variety of consequences that occur in high school. If one objective of BE is to promote the orderly progress of students through the school system and thus to improve life chances, then the grade span of new research might extend through the secondary school years.

Analytic populations of students in high school will be very different than in grades K-3, and classrooms are likely to display quite different features. While children entering grade 1 ordinarily speak L1 but are illiterate in either language, nearly all students entering high school are literate. Immigrant students are usually monolingual, fluent and literate in L1. They would either be assigned to bilingual education

or ESL classes, or be allowed to sink or swim in monolingual English classrooms. Contrasting these two groups would be extremely interesting. It would also be interesting to look at the apparent consequences for students who are not now in bilingual education, but were in bilingual programs in earlier grades.

In fact, a whole taxonomy of student characteristics can be constructed and used analytically. This would include the factors (1) U.S. born vs.immigrants, (2) grade and age on entry to the U.S., (3) number of years in bilingual education, (4) listening, speaking, reading, and writing skills in L1 and L2. The societal characteristic of their environment could be part of the stratification of the sample. For example, Llanes (1979) notes effects of community acceptance of new immigrants and their sociopsychological adjustment.

2.2 Selection of Stratifiers

No matter what the analytic method (quantitative, qualitative, or a mixture), and no matter how the sample is actually drawn (purposive or probability sampling), the study sample must represent the patterns of variations in attributes that might affect the bilingual education process. Kennedy (1979; also see below) has specifically clarified the importance of sample variation for making generalizations from case studies. In order to insure adequate representativeness of variation in the sample universe, in an efficient manner, the sample frame should be stratified. Although it is most desirable to stratify on bilingual education instructional features themselves, this information will not be available until a sample is actually selected. We must instead stratify on variables that are accessible, although they are only minimally likely to be associated with variations in bilingual education instructional features. Some potential stratifiers, to be used in a stratified probability sample or stratified purposive sample, are as follows:

Geographic Area

Language(s)

Cultural Group

Title VII /Non Title VII

Sponsoring Agency Type

Total Enrollment

Age of Project

Program Type

Grades Served

Per Pupil Expenditures

Fraction LEP

Fraction of Language Minority Students who are new immigrants.

We might use these stratifiers even in student-oriented samples (as from the OCR survey) to ensure that a variety of educational contexts is included. It is clear that this list of stratifiers is far from independent, i.e., one could never stratify factorially. This problem can be dealt with in several different ways. For instance one can calculate correlations among stratifiers and cluster them. Then one would choose a representative stratifier from each cluster. Alternatively one could factor analyze the stratifiers and use the first three or four factors as independent stratifiers.

One clear cut redundancy involves the state/county/city, Language, and Cultural group. For instance, projects in Miami are likely to be Spanish in language and predominantly Cuban in cultural definition. For most sites it would be impossible to select projects across a wide range of languages/cultural groups although there are notable exceptions like San Francisco. One must combine site selection with the Language/Cultural Group Stratification. In order to do this wisely it is desireable to characterize the sample frame through a survey, or to rely on existing surveys like that maintained by OCR.

By limiting the list of stratifiers to those on the OCR survey file (e.g.state/county/city, Language Enrollment, Numbers of BE children, Numbers of NES/LES children) the OCR file can be used directly in selecting

a stratified sample. Unfortunately several language groups (e.g. Portuguese, French, Italian) can not be identified in this survey. This omission could be corrected by using the geographic indicators and collecting additional data from SEA's or LEA's. It may also be possible to append data from parochial schools. Overall, the OCR survey approach appears strongest, despite these faults.

2.3 Site Selection

The PSU (geographical area, LEA, or school) must be selected before a detailed plan is constructed, but some general considerations for site selection can be given now. These considerations include the following:

- Representation of Cultural/Ethnic/Language groups, and the assurance of adequate generalizability to subpopulations.
- Opportunities for special site level anthropological studies. The site itself plays a role in the research since it provides the community cultural context. By grouping projects within the same site, site level anthropological studies can be conducted and used to support project level analyses.
- Opportunities for subsampling districts or schools within the site. In particular adequate numbers of secondary and tertiary units, displaying adequate variation across sampling strata and across bilingual education instructional features, should be present. If possible, both Title VII and non Title VII projects should be present, at same other variation in project rationale and control.

Thus if geographical areas are selected as PSU's, then (1) the sites must fit into a national, coordinated, stratified design; (2) basic demographic characteristics of the sites should be surveyed; and (3) characterization of the sub-site population of districts or projects should be

conducted. Primary sources are the Title VII roster of projects, SEA and LEA data, the OCR survey file, and the U.S. Census data. The latter is now considerably flawed by aging. The 1980 census data--at least some early counts--may be available in time for the 1981-82 academic year.

2.4 Sample Sizes and Statistical Power

Despite large overall sample sizes, the AIR study (Danoff et al., 1977, 1978) was hampered by small analytic samples for any but global analyses. In particular, it was not possible to pick out a reasonably large analytic subset of LEP children. Each subset intended for quantitative analysis should be large enough separately to sustain meaningful statistical hypothesis testing. Qualitative studies, too, must deal with reasonable numbers of cases, in order to be generalizable. Instructional features studies certainly involve different units of analysis, e.g. the child, school, district, or teacher. Each sample size will be different and each must be adequately large. Furthermore, the sample size of projects for each culture (e.g., Navajo projects or Puerto Rican projects) to which we want to generalize must be adequate. This can be determined by using statistical power analysis, applied to each quantitative hypothesis to be tested.

When the sample size is very large, practically every effect will be statistically significant, even though it may be artifactual, or too small to be practically significant. Power analysis can help in interpreting such data. When the power to detect an effect of less than practically significant size becomes substantial, then analytic interest in statistical significance should be shifted to interest in the actual size of the effect.

Expectations about the size and direction of the relations of BE instructional features to child consequences are shaped by our knowledge of the effects found in previous studies. Although there do appear to exist some scattered positive findings, there are also many so-called

"null effects". Many studies that find "no significant" difference are wrongly interpreted as statistical proof that the null hypothesis is true, since practical significance and statistical significance are confused in the minds of many readers. This basic misunderstanding of the principles of statistical inference can be corrected only by the willingness of researchers to consider statistical power as well as statistical significance. Unfortunately, the principles of statistical power analysis (Cohen; 1977) are not generally known, but these techniques should be applied in any new quantitative analyses to be conducted.

In a recent article, Cohen and Hyman (1979) state that fully 70% of the studies published in the journals of the American Psychological Association, the American Educational Research Association, and the Council for Exceptional Children lack adequate statistical power. They propose four "game rules" involving the size of practically significant true effects and the choice of alpha (allowable first kind error). These rules are to be stated before analyses are performed and reported. The study report would list its game rules and state the associated statistical power. If this were done universally, it would be obvious that most studies following the quasiexperimental paradigm had little chance of demonstrating either positive or negative effects. We recommend that a scheme similar to that suggested by Cohen and Hyman be implemented in any quantitative BE studies to prevent misinterpretation of "null" findings.

Sample sizes required for reasonable assurance of the visibility of effects are larger than generally thought. A common definition of the minimum educatinally significant effect size is .25 of the child-to-child standard deviation of the outcome measure. The most common level of significance used in hypothesis testing is .05 of the time, i.e., we wish to reject the null hypothesis by chance alone no more than .05 of the time. The a priori probability that an effect of .25 SD is incorrectly declared null is .41 at a sample size of 100 (t-test at alpha = .05;

see Cohen (1977). Sample sizes less than 100 yield hypothesis tests of abysmally low statistical power and "null conclusions" have little meaning. This should always be made clear in reports dealing with small sample sizes.

2.5 Sampling Method

Some possibilities for actually selecting the sample include stratified purposive selection, stratified randomized selection, controlled selection, and simple random sampling. These techniques vary in credibility, generalizability, and purpose. In fact, since the study will probably consist of several components, several of these techniques may be used. For instance, a stratified probability sample based on a survey of children, schools, or projects would be statistically generalizable to the population from which it is drawn and would be highly relevant in determining public policies for that population. On the other hand one might purposively select projects known to display particular unique features for focused studies. Thus, the selection of the sampling method is dependent on the details of studies to be performed and cannot yet be determined.

Acceptance Rate Problems

A wide spectrum of BE projects and conventional schools should be included. Unfortunately, the refusal of LEA's, schools, or projects to participate is likely to be a problem, especially for non-Title VII centers, for several reasons: lack of direct dependence on federal funding through grants; distrust of the Federal Government; lack of cooperation from SEA officials; lack of any perceived advantage; suspicion that unfavorable evaluation will occur, and so on.

This pattern, or its analog, was encountered in the National Day Care Center Study, which also concentrated on the study of natural program variations. Of federally funded centers selected by AAI, nearly all were

willing to participate, but for-profit and other non-federally funded centers were difficult to enlist. They were essential to the design, however, so that considerable effort was expended to enlist them. Much of the success of that study derived from this effort.

The lesson for Instructional Features Studies is that advance site development work will be required to enlist schools, LEA's, and projects. This may involve early meetings with SEA's, meetings with CEIS, public meetings, and meetings with community representatives. The field verification to be conducted as part of this planning study will help to clarify the site development work that will be required.

2.6 Adjustability of Scale

Most studies of the scope of the proposed Instructional Features Study involve a national sample clustered or stratified in some way. The stratification is aimed at producing geographical or other forms of representativeness as well as to reduce error variances in variables to be assessed. Often, there is a need to produce results that are valid within distinct subpopulations. When the total budget is known the designer can construct plans that are reasonably satisfactory, i.e., they achieve adequate representativeness (a partly political motivation) as well as acceptable sampling errors globally within selected subpopulations. The problem comes when an unforeseen budget cut occurs and the original design must be modified. The easiest, most politically acceptable, way to modify the design is usually implemented, at great cost to the designer's original goals.

There are difficult political, credibility, and generalizability tradeoffs to make in any case of drastic rescaling. The costs in credibility, political acceptability, and generalizability that would be incurred should be anticipated as a design factor from the beginning. By making plans specific and flexible in advance, so that readjustments do not come as a surprise, the political costs can be minimized. Such a scaling plan should be provided for the presently proposed study.

Chapter 3
ANALYTIC STUDIES

A variety of analytic issues for bilingual education instructional features studies have been identified, through discussions with NIE and our consultants, examination of the literature, and our own experience with other educational research studies. The studies most influential on our thinking include the Instructional Dimensions Study (Leinhardt, 1978a, 1978b), the Beginning Teacher's Evaluation Study (BTES), and the American Institutes of Research Study on Title VII (Danoff et al., 1977, 1978). We present the issues without making definite recommendations, except in a few instances where the way is crystal clear. For instance, the need for in-classroom observation systems and integrated ethnographic and quantitative frameworks seems indisputable. The next paper in this series will present alternative designs.

We will consider a number of analytic issues, but we do not claim that our treatment is exhaustive. The issues to be considered include:

(1) merging quantitative and qualitative paradigms,
(2) diversity of cultural linguistic groups,
(3) language proficiency/dominance testing issues,
(4) analytic subsamples of students,
(5) the role of comparison groups,
(6) contextual effects,
(7) control for teacher/aide qualifications,
(8) definition of treatment,
(9) generalizability,
(10) longitudinal studies, and
(11) the time frame and measurement of change.

3.1 Merging Quantitative and Qualitative Paradigms

There is an accumulated disappointment in purely quantitative methods as they have been applied in educational evaluation research. This disappointment is evident in the reception given the AIR study on Spanish-English Title VII projects. In order to understand the widespread criticism of the AIR study, and the sharpness of the debate between its critics and its defenders, it is necessary to realize how deep the gulf is between quantitative and qualitative approaches. According to Reichardt and Cook (1979), researchers are opposed at the paradigmatic level. Table 1 presents an adaptation of Reichardt and Cook's characterization of the attributes of the qualitative and quantitative paradigms.

Many authors (e.g. Campbell, 1979; Llanes, 1976; Trend, 1979; Berliner and Tikunoff, 1976), as well as Reichardt and Cook, have called for merging qualitative and quantitative methodology. NIE, in issuing the RFP that led to this study, asked for an examination of ways to merge qualitative and quantitative information. Our examination of these issues has led us to concur with Reichardt and Cook that the problem is at the paradigmatic level.

Quantitative methodologists have typically been committed to their paradigm, whole cloth, unwilling or unable to compromise at any point. The AIR study is a classic example of an attempt to force-fit a study into the quantitive quasi-experimental framework. In order to merge methods - as many analysts and educational researchers are suggesting - the attributes of the paradigms should be examined attribute by attribute. The Reichardt-Cook definition of the two paradigms provides a useful structure on which to base the discussion of proposed new BE instructional features research. The organization of our discussion follows this structure but includes two further attributes. They involve (1) approaches to causal interpretation; and (2) small scale versus large scale data handling in terms of encoding, storage, retrieval, and integration.

Table 1: Attributes of the Qualitative and Quantitative Paradigms*

Qualitative Paradigm	Quantitative Paradigm
Advocates the use of qualitative methods.	Advocates the use of quantitative methods.
Phenomonologism; "concerned with understanding human behavior from the actor's own frame of reference."	Logical-positivism; "seeks the facts or causes of social phenomena with little regard for the subjective states of individuals."
Naturalistic and uncontrolled observation. Assumes a dynamic reality.	Obtrusive and controlled measurement. Assumes a stable reality.
Subjective.	Objective.
Close to the data; the "insider" perspective. Valid; "real," "rich," and "deep."	Removed from the data; the "outsider" perspective. Reliable; "hard," and replicable data.
Grounded, discovery-oriented, exploratory, expansionist, descriptive, and inductive.	Ungrounded, verification-oriented, confirmatory, reductionist, inferential and hypothetico-deductive.
Process-oriented.	Outcome-oriented.
Valid; "real," "rich," and "deep" data.	Reliable; "hard," and replicable data.
Ungeneralizable; small samples or single case studies.	Generalizable; large samples.

*This table is adapted from Reichardt and Cook (1979). Some attributes have been combined and others discarded in an attempt to reduce overlapping.

Phenomenology vs Logical-positivism

The logical positivist holds that all meaningful statements are either analytic or are verifiable by observation and experiment. Phenomenologists build understanding of behaviors and beliefs from observations close to the perspective of the actors. Ethnographers are quintessential phenomenologists who believe that descriptions of activities have been limited through the use of logical positivist instrumentation. Observation systems such as Stallings' or Leinhardt's are seen to be highly selective filters or screens that partially obscure the phenomena under observation, and thus record only a part of what actually goes on. Actually, both approaches have yielded useful results.

This was recognized by Berliner and Tikunoff (1976), who attempted in the BTES study to combine quantitative and qualitative methodologies. They wished to identify variables across which more and less effective classrooms vary without constraining the investigation by using any of the traditionally oriented observation instruments. To do this they developed a list of instructional dimensions by ethnographic means. Ten more and ten less effective teachers were identified by examining gain scores of tests addressed directly to two week Experimental Teaching Units (ETU's). Ethnographers observed these classes (without knowledge of which were effective and which not) and prepared prose ethnographic protocols describing classroom behaviors. The protocols were paired (more versus less effective) and analyzed for variables, concepts, or dimensions that described inter-classroom variations. The very large initial list was pared to 61 mostly nonoverlapping variables. These were then instrumented via rating scales, validated by reexamination of the original protocols. This technique combines phenomenological and logical-positivist approaches effectively. It begins with an open, atheoretical, observation that excludes little, and proceeds to scales that can be used in quantitative analyses.

Since the BTES ethnographic-based scales were validated only on the original sample of high/low effectiveness classrooms, the generalizability/particularity of the BTES scales. Particularity to the resesarch sample is an attribute of qualitative methods that is, not completely acceptable in a study whose purpose is, ultimately, to bear on policy decisions. Yet, the BTES scales could easily have been tested on a new sample of classrooms to determine the generalizability of the scales as legitimate classroom dimensions. Furthermore, the predictive validity of the scales could have been assessed by examining the effectivenss of classrooms that rated high and low on the scales.

The BTES approach is a particular example of using phenomenological approaches as a basis for development of quantitative instruments. Variations of this method can and should be used in bilingual education instructional features studies, but the development of scales should also include verification of generalizability and predictive validity.

The differences between bilingual and monolingual classrooms, i.e., the features that distinguish bilingual classrooms, can be investigated by an adaptation of the BTES technique. Bilingual and monolingual classrooms would be substituted for effective and less effective classrooms and the same technique applied. Then the validation of generalizability and predictive validity would be carried out. The result would be a set of valid scales that capture the main features that make bilingual education different from standard education. Of course, we are interested in classroom practices that are similar in monolingual and bilingual settings as well as those that are different.

Carrasco, Guzman, Erickson and Cazden (NIE-G-78-0099) are presently exploring the problem of quantifying ethnographic information in various forms by allowing two researchers trained in videotape analysis to identify, in agreement with each other, segments of routine classroom events as units of analysis, units which will allow analysis not only at the qualitative level, but also at the quantitative, since there will be more than three similar units identified for each researchable question. While the information gathering process was ethnographic, the researchers' approach to breaking down the videotapes (and accompanying field notes, teacher interviews, aide interviews, and static information on all classroom subjects) into units of analysis is not ethnographic

per se, but rather, a technique that can be applied only after a good bit of open ethnographic information has been gathered. The Carrasco et al. coding technique, which includes all the information collected, should be considered for use in the Bilingual Education instructional features studies.*

Thus, it appears that logical-positivist and phenomenological approaches can be combined and that they need not be identified (respectively) with quantitative and qualitative approaches. An integration of this kind is essential for the bilingual education instructional features studies.

Naturalistic and Uncontrolled Observations vs Obtrusive and Controlled Measurement

Quantitative educational researchers have never been able to implement a true experimental approach because of the difficulties inherent in implementing an educational "treatment," or in implementing random assignment of individuals to treatment and control groups. Much of the discussion in evaluation research journals centers on this problem and means to adjust for it. The result has been the "quasi-experimental" paradigm, i.e., the best fit of the experimental paradigm to the realities of educational data collection. Unfortunately, it is a force fit, and the advantages of the experimental paradigm are largely lost.

Opportunity to implement the quasi-experimental paradigm, let alone the experimental paradigm, is limited and perhaps impossible in bilingual education studies. The principal stumbling block is in finding sufficiently well matched comparison groups. This difficulty is well documented by Danoff et al. (1977, 1978). It is a profound problem and we see no easy solution. There is no opportunity to do valid effects evaluations with properly matched control groups, and they should not be attempted.

The diverse universe of Bilingual Education as it is, provides a naturalistic, phenomenological basis for study. We do not believe, however, that conventional qualitative methods can be used to manipulate and synthesize the data. Except for exploratory analyses, the need for generalizability

*This paragraph was contributed by Robert Carrasco.

and large-scale data handling preclude this. To solve this problem, better techniques must be developed to capture qualitative variation in coded form, to deal efficiently with large amounts of data, and yet not to constrain analyses too severely. To do this there must be a substantial (but, we believe, feasible) advance in the state of the art.

As the BTES study suggests, a melding of observational data into a controlled measurement pattern proceeds to assume that what is a stable reality which can be perceived only through the controlled observation of the dynamic reality. It is in this realm that one may begin to make assumptions about causality. Yet causality, itself a stable reality, continues to modify and change the dynamic reality which the ethnographer observes. The strength of this approach may be tested via the generalizability theory outlined by Medley and Mitzel (1963) and Cronbach et al. (1972). Generalizability theory is an extension of classical test theory to the case where there are multiple sources of variation. In applying it to observation processes, temporal variation is one of the most important of these sources.

Subjective versus Objective Approaches

Both quantitative and qualitative methods are subjective. It has been the author's experience with quantitative evaluation studies that his own subjectivity entered substantially in selecting measures, constructing linear models and, especially, in the interpretation and integration of analytic results. Nevertheless, the impact of the subjectivity of the researcher should be minimized in studies that bear ultimately on policy decisions. If research is from a definite political, educational, or social perspective, then that perspective ought to be identified and well understood by those who use the research.

The biggest problem with case studies and ethnographic studies, from the point of view of the reader who wants to form his or her own judgments, is that comprehensive summaries of the primary data cannot usually be

directly presented in research papers. The data can be exhibited only in forms where subjective selection, synthesis, and integration have already entered. The reader must trust the researcher more than in quantitative studies. The audience for the proposed instructional features studies is too wide, too varied, and too important for policy, to rely on this kind of trust. The scale and importance of the studies demands a maximum of objectivity, and the visibility of data and analytic methods to consumers of the study.

One of the contributors to this report, Robert Carrasco, disagrees with the thoughts expressed in the paragraph above, and the general charge that classroom ethnography is partly subjective. The following quotation expresses his point of view:

> While the anthropologist or the ethnographer studies some previously unknown aspect of the culture, the classroom ethnographer already has some idea of the workings of the classroom. It is very likely that the classroom ethnographer has not only experienced this type of schooling in this country, but may also have been a teacher, an administrator, or may have extensive knowledge of the schooling system. Some classroom ethnographers also share in ethnicity of the subjects being studied (For example, Carrasco is studying Chicanos in bilingual classrooms. Carrasco is also Chicano. Carrasco has taught and has been educated in the Chicano community and has studied the education system before doing research in classrooms.) Therefore, the more knowledge the researcher brings out about the type of schooling and the community from which the school members come, the more accurate will be the descriptive information gathered in the field, and the more objective the descriptions will be in making sense of particular salient features in the classroom. One may pose this question: "Who can make more sense of what happens in the classroom? The person who has gone through or has some experience in that schooling, or the person who has no idea of what to expect in the classroom?"
>
> It seems to me that the classroom ethnographer is more objective in his/her description because he/she knows more about what goes on in the classroom. And with this previous knowledge, the researcher may be able to identify quickly and more efficiently (time-wise) the salient aspects of education in the classroom that have not been previously surfaced nor described in educational literature.
>
> Finally, classroom ethnographic procedures require that the teacher, aide, and even children be collaborating researchers; to

serve not only as subjects, but also as informers and as interpreters and provide analysis of certain features to be studied. In fact, this is the strength of this type of research. Classroom ethnographers make use of ETIC and EMIC distinctions and interpretations if the claim of certain findings are to be valid. Hymes (1979) and Hsu (1979) both agree that these ethnographers should be of the same ethnic/cultural background as those subjects in situations under study if the research is to be valid. If the ethnography is done correctly, the reader can trust the research. Qualitative research in general is not all the same. Ethnography is only one of many other approaches that fall under this umbrella, "Qualitative Research." Some of these approaches are indeed very subjective and very unreliable. Some raise more questions than they answer! Ethnography, particularly classroom ethnography, seems to fall into a more valid and more reliable arena, since the descriptions and analysis are made in collaboration with participants in the scenes studied. An ethnographic analysis validated by the teacher, it seems to me, leads the reader/listener to "trust" that which is presented as valid. Unlike quantitative researchers, the ethnographer is able to share with the reader/listener the RAW DATA in the form of field notes, interviews, videotapes, and what is becoming even more crucial in conference presentations, to have one of the participants present--mainly the teachr. If there is to be distrust, it should be at the raw data level, data that are never shared in papers and presentations by researchers whose only means for research findings are quantitatively based.

Nevertheless, AAI believes the problem of objectivity is intrinsic in the general approach we are suggesting. We favor discarding the quasi-experimental paradigm and using ethnographic methods. Yet these methods are seemingly subjective and hence problematic for policy studies. Obviously, the resolution of the difficulty will not be easy. Some possibilities are as follows.

- The studies might be broken up into substudies that are specifically intended to reflect particular points of view. The results of these studies might then be debated and, to a degree, data re-examined by their critics. Particular areas of disagreement can be identified and resolved through modified adversary processes. Synthesis and integration would follow. This is an approach similar to that adopted by the Center for Applied Linguistics in the process that led to the bilingual education Perspectives Series.

- Data should be stored in the most retrievable, primary forms possible for potential reexamination by adversaries. This would argue for the use of

videotape observation techniques rather than in-classroom coding. Videotape records are permanent, primary, and accessible at any time.

- In a large-scale study employing ethnographic methods, there will be necessarily many researchers involved. These researchers can be organized in a way that minimizes the influence of any particular person, e.g., by training or rotation. The use of videotape as primary data would facilitate checks by one analyst on another and minimize subjectivity of the primary observation. The biographies of researches might be included in reports in order to allow the reader to identify their perspectives.

"Insider" vs. "Outsider" Perspectives

The ethnographer typically aims for an "insider perspective." Mehan, for instance, customarily involves "participant researchers," i.e., classroom teachers who also participate as full-fledged members of the research team. Participant researchers are paid fees and are consulted regularly in the interpretation of videotaped behavior. The advantage to the insider approach is that it helps in interpreting the significance, causal connections, and sequencing of behaviors, by providing another perspective. Gross errors of interpretation can be avoided.

Unfortunately, the insider perspective may become too subjective, and the participant researchers approach may be too expensive and difficult to administer for large-scale research. Nonetheless, it is possible, as the following example shows.

In the National Day Care Study (Ruopp et al., 1979), each of the 64 day care centers participating in the study was granted funds for an on-site "center secretary" of their own choosing. The center secretary helped the center keep its books, sometimes assisted in classrooms, but was also responsible for on-site data collection of several kinds. She was a participant in the Day Care process as well as as a reporter. Her reports therefore reflected an insider perspective. There were other advantages as well.

First, the center received something tangible for its participation in the study. Second, it helped to build mutual trust between the study mananagement and day care study management. The only real disadvantage to this method was its cost. Setting up a system like this for BE instructional features research is well worth considering. For instance, secretaries could be provided to BE administrative offices and made responsible for data collected (across several schools).

Another alternative that should be explored is to employ community persons or teachers as on-site researchers.

If this elaborate a system for insider participation is not implemented, it is essential that practitioners be consulted at several points in the study. NIE realized this when it indicated the desirability of consulting practitioners in verifying BE definition to be developed in the course of this planning study. We would recommend that as data collection systems are developed, practitioners are again invited to express their views and to contribute to the development process. Further, when behavior is interpreted, practitioners should again be consulted.

Exploratory vs Confirmatory Analyses

AAI and NIE personnel, AAI consultants and other persons we have contacted all agree that much of the substance of Bilingual Education may be lost if only confirmatory analayses are conducted. It seems clear that initial exploratory analyses should give way to confirmatory as appropriate hypothetical stuctures emerge. Federal policy should ideally have confirmed knowledge directly applicable to formulation of laws and regulations. Our design problem is to find the appropriate ways of implementing initial exploratory and later confirmatory analyses efficiently, making use of the strengths of each approach.

Exploratory analysis can be used in quantitative studies as well as qualitative. In the newly evolved field of "Exploratory Data Analysis"

(Tukey, 1977), for instance, analysis begins with a search for hypotheses that are suggested by patterns in the data. A number of new tabular and graphic techniques (box plots, stem and leaf displays, etc.) are used to detect relationships that can lead to hypotheses. Some of these techniques were used by the author in the National Day Care Study (Ruopp et al., 1978). We suggest that they be used in BE instructional features studies.

These techniques are designed to make results less sensitive to atypical small subsets of the data, and on assumptions about the underlying statistical distributions.

A further suggestion is that for quantitative analyses, new techniques in robust statistics (see, e.g., Mosteller and Tukey, 1977) be used extensively. Although these techniques are often expensive to implement, they help to liberate inferential processes from untenable assumptions, and the credibility of results is considerably enhanced.

Ethnographic exploratory techniques like that of Berliner and Tikunoff (1976), discussed earlier, are also useful in avoiding confirmatory techniques in early stages of analysis. These techniques still lead to quantitatively analyzable data and ultimately to confirmatory analyses, and thus combine qualitative and quantitative aspects in the way that appears required for the proposed new studies.

Note that these two techniques are in a nearly symmetric relation. The first technique (EDA) applies exploratory, qualitative methods to quantitative data, while the second applies quantitative techniques to exploratory, qualitative data. In either case, however, the analysis must eventually progress to confirmatory stages. Exploratory analyses would also be used to investigate specific mysteries that arise in later stages.

Process Oriented vs Outcome Oriented Analyses

Although orientation towards product rather than process has been characteristic of quantitative approaches in the past, this is no longer

true. The development of observational instrumentation to measure the dimensions of process is one of the most important recent movements in educational research. Early, dissappointing, results of process-product studies seem now to be a result of insufficiently developed methodology (with regard to the validity, reliability, and stability of observation-based measures) and of concentration on areas that are instrinsically difficult to observe (behavior patterns, competence, teacher characteristics). Both the Instructional Dimensions Study and this BTES study used generalizability theory appropriately to approach the validity/ reliability/ stability issues and concentrated on global patterns of time use in the classrooms. These studies found robust positive relationships between process and outcome. Since Danoff et al. (1977, 1978) also found relationships in bilingual education between outcome, time use, and grouping patterns, even with questionnaire-based process data, a properly designed process-product study seems likely to uncover strong relations.

There is no longer any argument, in other words, that process is important in quantitative as well as qualitative studies. We certainly are breaking no new ground by recommending it for the proposed new studies In fact, the major instrumentation of the proposed studies should be process observation instruments that integrate enthnographic, sociolinguistic, anthropological, and psychometric perspectives (Llanes et al., 1976).

Generalizability

Qualitative analyses have customarily involved very few cases, investigated deeply; quantitative studies have involved more cases but often with narrow or shallow measures. The results of qualitative studies have, accordingly, been attacked as ungeneralizable. In fact, there exists no substantial theory even to investigate the generalizability of qualitative analyses. We agree to a large extent with these criticisms. The object of the policy maker is to apply knowledge to new populations,

new sites, and new circumstances. The knowledge he or she applies should be generalizable to the domains where the policy is to be applied.

It is the large scale of the proposed new studies that affords a solution to the problem of providing generalizable new results that are not as narrow, or shallow, as those that typically emerge from quantitative studies. The large scale affords us the opportunity to include many cases in the analysis and to select these cases so that the domain to which the results generalize is clear. This clarity is further promoted by connecting most of the different studies and substudies by linking them together on a common sample. They would thus be generalizable to the same clearly specified domains. It may be desireable, however, to supplement the generalizable part of the design by case studies conducted in nongeneralizable population or project types.

A recent article by Mary M. Kennedy (1979) discusses problems in generalizing from single case studies. She discusses the relationship between the general case and the special case in program evaluation and points out that it is difficult to aggregate one assortment of special cases into a meaningful general case. There are no generally accepted rules for drawing causation and generalization inferences from case study data. Kennedy proposes criteria that might be the basis for such rules. These include the sample attributes (1) wide range of attributes across sample cases, (2) many common attributes between sample case(s) and the population of interest, (3) few unique attributes in the sample case(s), and (4) relevance of attributes. For bilingual education instructional features studies, we would consider context as one of the most important of these attributes. These are qualitative generalizations of the criteria for generalization from quantitative data, where a theory of generalizability does exist. The attributes of a good sample for qualitative generalization are therefore similar to those for quantitative generalization. At the treatment level, Kennedy suggests (1) wide range of treatment attributes across replications, (2) common patterns of treatment outcomes across sample cases, and (3) common treatment functions across

cases. Again, these generalize desireable attributes for good quantitative studies where the treatment is variable as actually implemented.

The Kennedy approach, in short, helps to bridge the gap between quantitative and qualitative approaches to generalizability by concentrating on the patterns of variation in the sample of cases and of treatments.

Making Causality Interpretations

Most large quantitative studies are correlational and associative in nature, and, as is well known, cannot address problems concerning causal relations without external hypotheses. Retrospective (versus prospective, experimental) studies cannot yield strong causality results by any analytic methods. This is true even of static path analysis studies (involving causal diagrams) as well as of conventional correlational studies. It is conceptually possible to address problems of causal direction among variables in panel data studies over several points in time, or by multiple time series methods. But in fact, results of the panel data studies have been very weak (because of lack of statistical power). Time series methods have hardly been tried because of their expense and obtrusiveness.

Panel data methods should be considered for purely quantitative parts of the new bilingual education instructional features studies, but that the sample design take into account the statistical power available to see causal relations when they do exist. If definitive results are obtained, they could be very exciting. For instance, one might detect causal interrelations among L1, L2, and general cognitive development, or among affective, cognitive, and behavioral measures.

Multivariate time series designs would be useful in understanding these relations. Some of the Cummins (1979) hypotheses can be explicitly tested via these methods. One hypothesis is that there is a threshold level of proficiency in L1 beyond which instruction in L2 skills becomes much more effective. Since this hypothesis is so important for design of

bilingual programs, research on it is of exceptional interest both to policy makers and to program personnel. The direct way to see causal/ relations, however, is through micro-ethnographic techniques. These serve to establish the links between concurrent and succeeding events in the classroom, which can then be used in correlational analyses. This approach cannot illuminate the relations between broad developmental domains, however, but merely between neighboring events in established contexts. In other words, both quantitative and qualitative methods allow us to deal with issues of causality, but in different spheres. Therefore, quantitative and qualitative methods should both be used.

The Problem of Scale in Ethnographic Research

Ethnographic research has been used principally in smaller scale studies (e.g., the eight-site Headstart study currently being conducted by Juarez and Associates for ACYF, or several of the projects of the Laboratory for Comparative Human Cognition, University of California at San Diego (UCSD), or the work of Erickson, Cazden, Vera, and Carrasco). Even in such studies the quantity of prose information becomes vast and must be condensed and integrated. Juarez and Associates appears to be at the upper limit of what can be done without reduction to numerical codes--they use an elaborate categorical indexing system. In larger studies it will be necessary to develop and use more efficient coding languages than ordinary sociological English. The language could be syntactical, with a wide variety of efficient symbols. Developed further, such a language becomes a quantitative coding system, grounded on more open initial development.

There is a large community of ethnographic workers in such institutions as Harvard, Michigan State, UCSD and others. The field is in developmental flux, pushing closer to techniques amenable to large-scale studies. The best hope for an efficient system for instructional features studies lies in extending known workable systems to adapt them to larger amounts of information. Robert Carrasco has developed a computer-based ethnographic storage and retrieval system for his project with Erickson and Cazden. His system may be an excellent starting point for new development.

Other approaches by Rodriguez-Brown, Walcer, and Shultz (1976); Burck, Shultz, and Rodriguez-Brown (1977); and Walcer and Rodriguez-Brown (1978) illustrate the potential of even very simple coding schemes. These variables included (1) the amount of time in Spanish and English usage by target children in teacher-child interactions, (2) the amount of Spanish used in peer-peer interactions, (3) changes in language use by context (bilingual versus regular classroom). These quantitative analyses were supported by qualitative analyses of time use by activity, differences from person to person, purposes of messages, qualitative analyses of peer-talk in English and bilingual classroom.

The "bottom line" for BE instructional features studies is that improved coding and data handling techniques must be developed and embedded within a computer-based retrieval system. The system can probably be developed from existing systems more suitable for smaller-scale research, and with a narrower analytic point of view. The development of a new system will require time to develop, shake down, and install, and it could not be in place during the 1980-81 school year. We discuss this problem later.

3.2 Diversity of Cultural/Linguistic Groups

The only previous large scale bilingual education evaluation study (Danoff et al.,1978) concentrated on Title VII Spanish-English bilingual projects. The bulk of smaller studies in the United States also involves this linguistic group. The emphasis of the literature has, in other words, reflected the pattern of Title VII funding, which is dominated by Spanish-English projects. Tables 2-3 (from RFP 79-100, U.S.O.E., 1979) exhibit this pattern. Of one-language Title VII projects, 370 (79%) are Spanish while the remaining 97 projects involve 45 other languages. Seven languages of the 46 account for 90% of the projects. Only Spanish, Navajo, and French involve sufficient numbers of Title VII projects to afford meaningful within-language statistical generalizations. Many languages could be included in studies that make cross-language comparisons of instructional features.

Table 2: Number of Languages Addressed by Individual Projects During 1978-79 Academic Year

Number of Languages	Number of Projects	Proportion
1	467	.824
2	56	.099
3	21	.037
4	9	.016
5	9	.016
6	3	.005
7	2	.004
Sum	567	1.000

Inclusion of several language groups in a major BE study involves technical problems because of the diversity of these groups. No rational study plan can lump these diverse languages and cultures in analyses, although parallel techniques can be used. Differences requiring special attention include the following:

Table 3: Languages Addressed by Individual Projects During 1978-79 Academic Year

Languages	Number of Projects
Spanish	370
Navajo	15
French	
Portuguese	
Italian	6
Crow	4
Mohawk	3
Dakota	2
French Canadian	2
Tewa	2
Chinese	2
Yup' ik	2
Inupiaq	2
Papago	2
Vietnamese	2
Arabic	2
Choctaw	2
Samoan	1
Aleut	1
Chamorro	1
Kusaian	1
Gwich' in	1
Ponapean	1
Marshallese	1
Athabascan	1
Trukese	1
Palauan	1
Pima	1
Havasupai	1
Piute	1
Greek	1
Delaponke	1
Ilocano	1
Passamaquoddy	1
Yiddish	1
Armenia	1
Ojibwe	1
Northern Cheyenne	1

Table 3 (Cont.)

Languages	Number of Projects
Keresian	1
Apache	1
Japanese	1
Haitian	1
Menominee	1
Central Yup'ik	1
Portuguese, Spanish	10
Italian, Spanish	8
Chinese, Spanish	4
French, Spanish	3
Filipino, Spanish	3
Spanish, Vietnamese	2
Tewa, Spanish	1
Korean, Spanish	1
Haitian, Spanish	1
Spanish, Yiddish	1
Italian, Japanese	1
Chinese, Korean	1
French, Haitian	1
Korean, Vietnamese	1
Creek, Seminole	1
Russian, Spanish	1
Arabic, Spanish	1
Pennsylvania Dutch, Spanish	1
Carolinian, Chamorra	1
Dakota, Sioux	1
Punjabi, Spanish	1
Navajo, Spanish, Ute	1
Ute, Spanish	1
Miccosukee, Spanish	1
French, Hungarian	1
Haulapai, Havasupai	1
Spanish, Tagalog	1
Ilocano, Spanish	1
Armenian, Spanish	1
Navajo, Spanish	1

Table 3 (Cont.)

Languages	Number of Projects
French, Haitian, Spanish	2
Korean, Spanish, Vietnamese	2
Spanish, Ilocano, Tagalog	2
Ulithina, Woleian, Yapese	1
Tewa, Spanish, Keresian	1
Algonquin, French, Spanish	1
Italian, Portuguese, Spanish	1
Arabic, Greek, Spanish	1
French, Italian, Spanish	1
Greek, Spanish, Vietnamese	1
Chinese, Italian, Spanish	1
French, Spanish, Yiddish	1
Spanish, Tagalog, Vietnamese	1
Greek, Korean, Spanish	1
Korean, Cambodian, Vietnamese	1
Spanish, Korean, Vietnamese	1
Chinese, Filipino, Spanish	1
Hebrew, Italian, Russian, Spanish	1
Greek, Yiddish, Italian, Spanish	1
French, Haitian, Italian, Spanish	1
Chinese, Greek, Haitian, Spanish	1
Cambodian, Chinese, Japanese, Korean	1
Cambodian, Chinese, Tagalog, Vietnamese	1
Chinese, Filipino, Spanish, Tagalog	1
Japanese, Korean, Filipino, Spanish	1
Chinese, Spanish, Japanese, Korean	1
French, Greek, Haitian, Portuguese, Spanish	1
Chinese, French, Greek, Italian, Spanish	1
French, Hebrew, Russian, Spanish, Yiddish	1
Vietnamese, Chinese, Japanese, Spanish, Thai	1
Chinese, French, Spanish, Japanese, Persian	1
Chinese, Korean, German, Russian, Vietnamese	1
Chinese, Korean, Spanish, Samoan, Tagalog	1
Chinese, Indo-Chinese, Portuguese, Spanish, Tagalog	1
Chinese, French, Greek, Italian, Spanish	1
Chinese, Greek, Italian, Korean, Spanish, Vietnamese	1
French, Korean, Japanese, Thai, Yup'ik, Spanish	1
Chinese, Filipino, Japanese, Korean, Samoan, Spanish	1
Chinese, Japanese, Korean, Filipino, Tagalog, Vietnamese, Laotian	1
Chinese, French, Italian, Korean, Portuguese, Spanish, Vietnamese	1
Sum	567

Now, however, a demand has built up in the bilingual education community for inclusion of non-Spanish projects in new research. Although the universe of Title VII projects does not afford an adequate number of projects for research in most languages, the expanded BE sample frame (including non-Title VII BE projects) should be considerably better. For instance, the design should include an adequate pool of Chinese projects, but in Title VII there are only two one-language and 17 multiple-language projects involving Chinese. This argues for expansion of the sample frame.

(1) Different needs and problems exist in the academic domain. Academic performance of Chinese students is, for instance, not markedly different from that of middle class whites; but academic performance of Hispanic background students is below that of middle class whites. Newer immigrant groups like the South East Asians present quite different pictures from the Chinese because of the independence of social class/income of parents as a variable.

(2) Home cultures are different in their support of the schools and in their involvement with schools as well as their developmental interests and activities. Certain communities are highly motivated politically and others are not. Any study of the community and home cultures and school-community links must be done separately by language group.

(3) The cultures are different in their aspirations for their children and the community, and therefore what they want from bilingual education. For example, certain Japanese enclaves do not ask for bicultural approaches within the school system but, rather, for speedy assimilation. Other cultures are confident that family and neighborhood institutions can carry much of the burden of maintenance of the native culture and language. The ESL approach may therefore be quite appropriate for one community, but inappropriate and unpopular among others who desire more cultural support from the public school system.

(4) Most importantly, there should exist valid and reliable language proficiency and academic proficiency measures for each culture that is

included in the main analyses. Language/culture groups for which such measures do not exist should not be excluded entirely, but would require special approaches, e.g., case studies. In deciding on the approach for any specific language group, it will be necessary to survey the measures available. This issue is discussed below.

Obviously, not all of the 70 languages involved in bilingual education projects can be included in the next round of bilingual education studies. It seems to us that the large scale components of new studies should not try to cover the whole spectrum of languages, but only, perhaps, three or four. The Spanish language should be differentiated by geography as well as other context variables. No previous large study has done this. Inclusion of Native American groups is essential, and the technical tools exist for linguistic measurements in Navajo. French, Chinese, and Portuguese projects should also be considered for inclusion. The Southeast Asian languages, particularly Vietnamese, should be included as an example of a one-time-only immigration wave. One could extend tht study to adults and high school students as well as primary school students. This language group would require the use of case study methods.

The technical problems involved with including several languages should not be minimized. It may be wise to separate those language/culture groups that can support generalizable analyses and those that can not. A meaningful way to organize and divide new studies would be to include the generalizable groups, with adequate language proficiency measures, in a large integrated study and to sponsor case studies in languages that can not be meaningfully integrated into a large study. These could be conducted by different research organizations, since little integration of the studies will be required. Comparison across languages is in itself interesting.

NIE has already begun to address some of these issues in its sponsorship of three important ethnographic/sociolinguistic studies for the native American, Asian-American, and Spanish cultures. At the time of this writing, only one of the contracts has been awarded. These studies may include such

elements as surveys of language speakers, description of dialects, identification of linguistic features, study of context, and assessment of attitudes towards languages and dialects. Methods may involve observation of communication, determination of participant structures, and assessment of language skills. It seems likely that these new studies will contribute substantially to an overall approach to the features study--i.e., their results can and should be factored into the actual features study. They should also figure prominently in producing alternative designs for the study. Accordingly, AAI plans to study each of the winning proposals, and to speak to the contractors, and to use this information in constructing designs.

3.3 Linguistic Proficiency Issues

The need for measures of proficiency in both English and the student's native language appears compelling, based on conversations with researchers and examination of the results of major studies (e.g., Ramirez et al.,1976). The need for valid proficiency measures exists in nearly every bilingual education study because it is relevant in so many ways, e.g.:

(1) valid classification of students into linguistic proficiency subgroups for purposes of analysis;

(2) determination of the distribution of linguistic proficiencies in classrooms as an instructional feature *per se*;

(3) determination of the entry-exit system operationalized by the school system;

(4) distinguishing differential (towards linguistic groups) interaction patterns by teachers;

(5) determination of relative language proficiency (L1 vs. L2);

(6) determination of the longer-term development of proficiency in both languages;

(7) consideration of the appropriateness of curriculum, classroom subgroupings, pull-out schedules, etc.;

(8) study of students' participation in classroom process as a function of language proficiency; and

(9) control for entry proficiency in language for assessment of student progress.

The definition of language dominance is not always clear in the literature, and is confused to some degree both with relative language proficiency. In this study, an important issue is whether dominance should be determined in addition to relative proficiency in L1 and L2. One can conceive of a continuum running from L1 and L2 for each of the four skill areas (listening, speaking, reading, writing), for each of the domains the speaker encounters (e.g., the classroom, playground, street, and home). Language dominance can be thought of as the set of loci on these continua. As such, dominance is not a single variable, but a complicated collection of context related variables. Because it is so complex, and few valid measures exist, it appears best to abandon the concept of dominance in the instructional features study. Note that in California the state originally mandated dominance measurement, and abandoned it as essentially unmeasurable. Proficiency measurements were substituted.

The difficulties that have been encountered by previous studies did not arise because the researchers were not aware of the need to measure proficiency. Rather, the difficulty has been efficient, valid, reliable measurement at acceptable cost. We think that administration of proficiency measures in all grades is essential, even though it may be very expensive (especially in grades K-1), because of the many uses for proficiency measures. Teacher opinions and child interviews do not yield valid proficiency scales, and their use has been attacked in previous studies.

3.4 Analytic Subsamples of Children

In no other case in large scale educational research is the identification of analytic subsamples of children so important. This is true first of all because the influence of status variables like entering language proficiency is so strong and complex that it cannot be adjusted away by covariable (ANCOVA) analytic strategies. Second, the outcome measures of interest and relevance for LEP students are different from those of English Proficient students. If the bilingual education instructional features studies are to influence federal legislation it is necessary to focus analyses on the Title VII target group, i.e., LEP students. In the AIR study

(Danoff et al., 1977, 1978), LEP students could not be separated for independent study, since only one-fourth of the Title VII sample and only 5% of the comparison group was LEP. This inablity to sort out different Bilingual Education consumer groups (non-English dominant LEP, balanced (proficient) bilingual linguistic minority, English dominant linguistic minority, and English dominant majority students) decreased its usefulness for public policy. For these purposes, dominance might be measured in terms of relative proficiency in L1 and L2 (Zirkel, 1974). The question of comparable scales then arises, however. Clearly, the concept of dominance is problematic. In designing the sample an adequate analytic sample of each significant consumer group must be assured.

3.5 The Role of Comparison Groups

One objective of the proposed studies is to relate the qualitative presence or absence of instructional features, or quantitative variations of specific features, to consequences for children. Features of specific projects would be assessed through observation or direct measurement rather than only through the sponsor's intentions. Since the Instructional Features Study is not intended to evaluate the effectiveness of bilingual education generically, or even of specific "whole models," identification of well-modeled comparison groups is not required. In any case, it is crucial to consider the intervening effects of several other factors, including

(1) SES,
(2) language proficiency/dominance,
(3) linguistic use/communicative competence,
(4) academic performance,
(5) relations between the school and the community on bilingual education,
(6) patterns of language use in the home,
(7) cultural adaptation, and
(8) length of time in bilingual education.

The importance of these variables is demonstrated by analyses from the AIR study. The subsample of LEP children, for instance, was surpassed significantly in CTBS scores by the English proficient children. The influences of the home and community on school children are so profound in the language domain that linear ANCOVA adjustments can not possibly correct adequately for inter-group differences. Comparison groups with analytically meaningful numbers of LEP children do not exist in sufficient numbers to support a national evaluation study, although they may still be useful in instructional features research. This point was discussed above.

We discussed these problems with people in several other universities and research organizations, including Southwestern Educational Development Laboratory (SEDL), Juarez Associates, BABEL, San Diego State University, University of California at Los Angeles (UCLA), and University of California at San Diego (UCSD) and find near unanimity on the issue: it is infeasible to locate fully adequate comparison groups for a large scale study of bilingual education. It appears to us, then, that we should abandon any hope of locating these groups for the purpose of student impact analyses.

Non-evaluative Uses of Comparison Groups

While it is not feasible to identify comparison groups for traditional impact evaluation, they should still be included in the study for a variety of other purposes. The OCR survey gives us the means to locate LES/NES children who are not enrolled in BE programs.

Glaser and Strauss (1967) discuss the use of comparison groups for purposes of theory generation, i.e., in the qualitative, phenomenological mode rather than in the quantitative, logical positivist mode. While we would still need to identify non-BE groups that contain LES/NES children, there is no need for the precise matching required in the quantitative paradigm. The OCR sample frame would make it possible to identify comparison children, if they exist anywhere in sufficient numbers.

3.6 Contextual Effects

The context of the educational process is always important, but especially so for bilingual education. Community clubs and halls, churches, and parents do not involve themselves much with teaching mathematics, or other content areas, but they are extremely important for the teaching of language skills. It is even probable that most language learning occurs outside the school. Communicative competency is especially important outside the classroom and should be included as a consequence for students in bilingual education studies. The more immediate extension of the classroom--school buses, lunchrooms, the playground, the gym, etc.--is also of great importance and should be included in any survey of significant instructional features.

It is technically difficult to deal with this problem. A quantitative approach can not be considered, since it is not possible to make sufficiently stable and reliable measurements of the processes involved. It is, however, possible to collect census-type data and to make observations on the presence and quantity of certain known products of language development, including such things as mass media programs in the minority language, church and other social programs in the minority language, and the quantitative effect of proximity to large bodies of speakers of that minority language. However, for the most part, the process variables are the interesting ones, and these include such items as attitude toward the minority language. Ethnography and anthropology are the disciplines that must be applied.

In the chapter on sampling above, one alternative was to consider the linguistic minority student as a sampling unit. Such a student is in a bilingual learning situation, whether or not he or she is enrolled in a formal bilingual education project. From the student we can extend the analysis to identify the formal and informal supportive institutions. The quantity and quality of the processes that occur must be determined, and used to support a more detailed, more quantitative analysis of that part of education that actually occurs in school.

The objectives, attitudes, and motivations of different actors in this system should also be examined. In particular we should look at the pressures for assimilation and/or maintenance so that we can understand the sociolinguistic changes that are occuring in the student. Community attitudes effect changes in variables within both the formal (school) and non-formal (community-home) schooling. These attitudes are pluralistically defined. The predominant community view, is likely to affect variation in hypothesized characteristics of students, treatments and outcomes.

It would be very difficult to "control" analysis of the instructional process for these contextual effects. The best we can do is to identify them and conduct analyses only within domains where external support is similar. The case study of family and community systems and processes is then presented as the analytic content.

One specific use of comparison groups would be to assist in establishing the dimensions and variables that describe the differences between bilingual and monolingual classrooms. As discussed earlier, a modification of Berliner and Tikunoff's (1976) technique could be applied. Another use would be to provide a basis of comparison of children's linguistic behaviors across monolingual and bilingual classrooms.

3.7 Teacher/Aide Qualifications

Qualification for Teachers

The appropriate qualifications for teachers in bilingual education are not well understood. Being a native member of the native culture is an advantage but is not sufficient in itself (Light, 1974). Issues involve the relative importance of membership of the culture, knowledge of subject matter, knowledge of educational theory and practice, and the need for specific linguistic knowledge (including knowledge of the systematic nature and social/cultural origin of dialects and language variations.)

The bilingual-bicultural task force of the California State Department of Education conducted informal surveys in that state (Salamanca, 1974). It was determined that only 60%-65% of the BE teachers were bilingual and only 50% bilingual-biliterate. In contrast, nearly all bilingual education teacher-aides were bilingual and a substantial fraction are biliterate. Clearly, there is a shortage of qualified BE teachers and the slack is being taken up by bilingual education aides. In the next section we discuss relevant factors in aide qualifications.

While it may not be possible to generalize outcomes to differences in teacher/aide qualifications, it is a significant feature of bilingual education that instruction is conducted by a linguistically competent person. It is extremely desirable to test teachers for language proficiency, but this is probably not feasible if the study is to be acceptable to the schools involved. Beyond the issue of linguistic competence outlined earlier in this report, lies the effectiveness of the teacher in professional terms. Analyses of variance are insufficient to account for the factor of teacher qualifications in a study as complex as the one being currently planned. Thus, the issue of teacher qualifications must be solved through a simple categorization process, the variables of which might emerge from the significant feature definition itself (program-related), and/or the intended outcome measurement on the part of the formal (school) plan of action, and whether it be (1) in response to a mandate, (2) affirmational of community need, or (3) reformational within the cultural context, or (4) transformational into a linguistically poor context of schooling.

Qualifications for Aides

A large number of bilingual education classrooms team monolingual teachers with bilingual aides. Thus the entire first language content is provided by the aide, rather than the teacher, and her qualifications become of great importance. Often, teacher aides are not well qualified

for their jobs, particularly in content areas. When the teacher is monolingual, the lack of qualifications of bilingual aides can be crucial. Problems of teacher aide incompetence in monolingual settings are well documented. In bilingual education, the problem was documented (e.g.) by the Arizona Occupational Research Coordinating Unit (1976). They determined that bilingual teacher's aides are specifically unprepared to perform in the very areas for which they are employed. This document recommends and describes courses in the skills that appear to be most important within the bilingual classroom. One common problem has been the additional difficulty in delivering instruction in two languages, as compared with informal discourse in the community and home. In other cases aides may actually be qualified teachers from foreign countries who do not have U.S. certification.

The wide variations in qualifications, and their importance in the bilingual classroom, demonstrate a need to assess aides' background and qualifications as a major determinant of consequences for students. The fact that Danoff et al. (1977, 1978) found no relation between child outcomes and teacher/aide qualifications makes us pessimistic about finding effects in new studies, but this area should not be ignored in the proposed new studies. If language proficiency of teacher-aides can be measured, there would almost certainly be relations of these measures with consequences for children in classrooms where substantial amounts of teaching are conducted by aides. In any case, the modes of use of teacher-aides in the classroom are important instructional features in bilingual education.

3.8 Definition of Treatment

Perhaps the most serious flaw of the AIR study (Danoff et al, 1977, 1978) was that it considered the Title VII treatment as a single well-defined variable. This approach ignored the facts that (1) the intent of Title VII treatments was extremely variable, i.e., the theoretical whole models differed from each other substantially; and (2) the degree of implementation of the treatment, was also highly variable.

NIE was clearly aware of these problems when it issued the RFP for planning assistance for instructional features studies. By stressing instructional features rather than whole models they indicated preference for an approach that takes into account the multidimensional nature of variation within bilingual education and between monolingual and bilingual programs. The instructional features approach is an extension of traditional instructional dimensions approaches to include the classroom context and beyond quantitative measures. (See Planning Papers 1 and 2). In effect, the treatment in such an approach is accorded the complexity in analyses that it displays in reality.

3.9 Generalizability

The bilingual instructional features study will likely consist of a complex amalgam of large and small, quantitative and qualitative studies. In any of these cases it will be necessary to specify the universes or subuniverses for which the results are valid, i.e. to which they generalize. For large quantitative and quantitative/qualitative studies with probability sample data bases, a statistical theory of generalizability exists and should be a factor in design. Smaller case studies are not addressed by statistical generalizability theory but require, instead, qualitative, verbal analysis of applicability. Domains of generalization would include at least the following:

- classes of Bilingual Education approaches (specification of these classes is another problem),
- cultural/linguistic/geographic groups,
- language proficiency groups,
- community goals, and
- years of program operation.

The decision as to the generalizability of results to any particular population requires only accurate measurement of the appropriate traits and the application of statistical theory and common sense. Many studies claim generalizability beyond what can be supported rigorously through the use of elaborate verbal arguments.

It is not possible to restrict the instructional features study to generalizable sub-populations of language/culture groups without eliminating some of the most interesting and policy relevant populations. Several of the Native American tribes fall into this category because the total number of projects is not sufficient to support statistical analysis within the population. We cannot justify counting out these groups for this reason. Their inclusion in the study will almost certainly lead to valuable insights and the development of new, qualitative knowledge. These groups must be treated differently from more widely represented cultures, however. For instance quantitative and quantitative/qualitative conclusions might be based on larger groups and the results compared to quantitative and qualitative analyses of the smaller groups. I.e. the data from the smaller groups would be interpreted in the light of that from the larger.

3.10 The Time Frame and Measurement of Change

The RFP that led to the present planning study stated that major results of the study must be available during the period Winter 1981 to Spring 1982, in time for the ESEA reauthorization hearings. One way to insure this is to collect data during the 1980-81 academic year and to analyze it in the following year.

If that plan were adopted, the study would be badly constrained. Selection of tests and measures and the long process through FEDAC/CEIS, SEA's, and LEA's could not begin until a contract was awarded, which would be Summer 1980 at the earliest. It would be difficult to implement the first test administrations before December or January. This means that a pre-post design implemented in 1980-1981 could have only a very short test interval. No matter what precise approach is taken to the problem of measurement of change (change scores, standardized change scores, use of pretest as a predictor of the posttest score, residualized change scores, etc.), one predominant technical problem is evident (Cronbach and Furby, 1970) - the psychometric reliability of the change measure. This problem is most severe when there is a very short test interval. The psychometric

reliability of the change measure is often as low as .10 or .20. As a result, hypothesis tests on regression coefficients fail to reject the null hypothesis even when it is false.

This may be part of the explanation for the lack of significant effects in the AIR study. It will also be a problem in any new analyses that are conducted. There are two routes that can be taken to improve on the situation, but neither will make true changes visible with certainty. (1) First the test interval must be made as long as possible to allow as much new instruction-induced variance to enter as possible. (2) Second, we should use measures that are most likely to register change. The best tests (i.e., those most likely to show effects) are those tailored specifically to the intent and implementation of the curriculum. Usually, too little of the test content directly addresses the intervening instruction.

A second alternative is to devote the 1980-81 academic year to instrument development and implementation of systems, and to collect data in the 1981-82 school year. This would also constrain analyses badly, if any time is to be left for analysis in time for the 1982 hearings. Thus, a useful pre-post design doesn't appear to be feasible in that year either.

A third option is to test midway in the 1980-81 academic year and to repeat the measure one year later in the 1981-82 school year. This option is attractive in terms of the measurement of change. There might be considerable attrition, however, and possible discontinuity in the educational environment across grades. Nevertheless if change is to be measured--and this may not be required--this design is superior to the others considered. There are considerable problems in selecting the testing language and many other related problems that make academic performance measurement problematic.

It should be remembered that this is not the only bilingual education study being conducted for Congress. The instructional component study

being conducted by Development Associates for USOE is mandated to produce results within the legislative time table. This may make it feasible in the NIE study to concentrate more on the ultimate quality of the product for the larger bilingual education community. Freed from time constraints, a far stronger study could be implemented. Bilingual education may not be able to survive an endlessly accumulating mountain of hastily done research.

3.11 Longitudinal Studies

Language acquisition is a developmental process and should not be considered independent of general cognitive development. If language acquisition is to be studied meaningfully it should be in a longitudinal design. Other motivations are the need to investigate longer term socio-economic consequences, academic success in high school, and any other changes that can be assessed. Interviews would play a major role in such studies.

However, follow-up studies bring their own problems. One criticism of the AIR study was the short intertest interval (5 months from Fall 1975 to Spring 1976). To reply to this criticism, AIR identified a follow-up group to be tested in Fall 1976, a full year after the pretest. Because of funding problems only a subset of classrooms could be considered. These contained 277 Title VII and 71 comparison children in Grade 2. But there were only 44 children with complete data in sites with comparison children, and only about 15 of these children could have been LEP. Thus, the sample sizes used in follow-on analyses were too small to provide meaningful conclusions.

The need for longitudinal studies should be anticipated and an efficient roster file system installed in the first study year. This file would make tracking and follow up easier. If NIE decides on a two year approach to measurement of change, then the file would be valuable for shorter term analyses as well.

Tucker (1979) finds that the lack of longitudiunal studies is one of the most severe problems in obtaining hard empirical data about BE:

> Critical, empirical, and longitudinal assessments have been particularly lacking. The paucity of such studies can be attributed to a variety of sources such as the transience of potential researchers, the unwillingness or inability of funding agencies to commit funds for a multiyear or even decade long project,...and the unfortunate tendency for administrators to regard initial results as the major criterion for continuing or terminating a proposed lengthy project.

It would be an advantage indeed if, at this time, NIE could begin such a longitudinal study.

Attrition

Most so called longitudinal analyses performed are in fact "quasi-longitudinal," i.e., they compare cross-sectional averages grade by grade for each cohort. These analyses maximize sample size but are subject to attrition effects. Students who are lost from the sample often differ from those who remain in important ways. For instance, successful children may be systematically diverted to regular classrooms. A variety of similar biasing mechanisms exist. These are strong arguments for avoiding quasi-longitudinal analyses.

However, a longitudinal sample may represent only a small portion of the original sample, and the results of a longitudinal analysis may not be generalizable to the original population of interest. The adequacy of a longitudinal sample for supporting statistical analyses is easy to check using tables in Cohen (1977). But even when statistical power remains adequate, the generalizability issue may be problematic.

It is customary to check the equivalence of attrited and non-attrited groups via t-tests (or similar tests) involving language proficiency or any of a variety of demographic measures. These tests are usually not particularly useful except when equivalence is rejected. Failure to reject the null hypothesis can as often be blamed on lack of statistical power as on true absence of a biasing effect. We should accept such tests as evidence of no difference only when statistical power is demonstrably adequate, i.e., at least .90 for group differences of practically significant magnitude. This will not often be the case.

There is little that can be done to avoid these problems except to ask for very large initial samples and to track all children via a fairly sophisticated record keeping system.

Chapter 4
MEASUREMENT ISSUES

Here we consider some measurement issues arising out of analytic requirements for BE instructional features studies. Since we are not yet at the point of setting forth explicit study design alternatives (to occur in Planning Paper 4, to be submitted February 1, 1980), we do not yet know what measures are required, although the following domains might be included:

- o Classsroom Process;
- o Language Proficiency (L1, L2);
- o Family/Community Context;
- o Language Use/Communicative Competence;
- o Academic Performance; and
- o Affective Domain.

The need to consider such a broad domain of variables is justified in Planning Paper 1. There are obviously serious measurement issues in each area. The first two of these domains will be involved in any rational study plan, as indicated in Chapter 3. Because these areas are exceedingly complex, it is useful to consider some of the general issues involved in this report.

In Planning Paper 4, measures required for specific design alternatives will be identified, and the issues further developed. The feasibility of these measures will be considered in the field verification effort, to occur in February, 1980. Finally, specific alternative measures, and specification of information collection systems will be presented in the Interim Report (July 1, 1980). These will be based in part on surveys of available instruments that are currently being conducted.

4.1 Classroom Process

The discussions in Chapter 3 point the way to study designs that feature measurement of classroom process and its proximate extensions (school and schoolyard). Domains of process that could be addressed include

- o classroom time use by language and subject matter;
- o linguistic use patterns by teacher and students, differentiated by student characteristics;
- o teacher paralinguistic communication (Trager, 1959);
- o congruence between verbal and nonverbal teachers' behaviors;
- o participation structure;
- o classroom interaction (after Flanders);
- o school wide climate;
- o classroom space and materials;
- o program management;
- o features of the immediate extension of the classroom into the playground, school bus, lunch room, gym, etc.
- o teacher/aide qualifications and competencies; and
- o classroom organization.

These are somewhat overlapping, but they indicate the scope of measurements that would be included in any comprehensive instructional features study. This list of domains makes it obvious that measurement is a multidisciplinary problem calling for the skills of traditionally oriented classroom process researchers, macro-ethnographers, micro-ethnographers, and linguists. Further, no one measurement technique is adequate. Interviews, questionnaires, teacher ratings, videotape direct observation, in classroom coding and checklists may all be required. The technical problems involved are substantial.

First, if process measures are to enter meaningfully into an overall study of instructional features, then the information from diverse disciplines and data collection techniques must be integrated, both conceptually and technically.

Second, if the data are to be used in any but small-scale analyses, they must be efficiently coded and stored in a computer-based data system. A new software system might be required.

Third, methods must be found to access and integrate data via an efficient retrieval system and suitable analytic software.

The latter two problems do not apply to small scale (perhaps special purpose) substudies. These studies will undoubtedly be conducted, but should pose no new special problems. Larger scale analyses, will require tangible improvements in the present state-of-the-art. The study design will therefore have to include substantial time for development, implementation, and installation of observational data systems.

4.1.1 Alternative Measurement Techniques

In the next pages, we consider the strengths and weaknesses of a variety of measurement and information collection methods.

Questionnaire and Interview Data

Interviews and questionnaires are indispensible in obtaining data on teacher/aide qualifications and attitudinal data. They have also been used to collect information on interaction processes, but much less successfully. It can be argued that in a large scale study, intended to have a direct influence on legislation, direct observation is the only valid way to assess much of the interaction process. Nevertheless, interview and questionnaire techniques have their place in these areas. The AIR study showed the usefulness of these techniques in bilingual education instructional research.

In the AIR study, process data were obtained from interviews with classroom teachers, team teachers, special ESL teachers, special Spanish language teachers, and special mathematics teachers. The data were integrated to calculate the 25 classroom level process variables listed in Table 4 (from Danoff et al., 1978). The analytic strategy was apparently to determine time-use in the classroom across a number of relevant categories. Their approach was justified by the potency of such variables as academic learning time in the BTES study, and general predictive power of time use variables. In fact, consistent relations between children's test performance and the extent of grouping/regrouping were discovered. Perhaps some additional relations between process and outcome could have been detected if measurements had been by direct observation, provided stable measures can be constructed. One advantage of direct observation is that the reliability of measures can be computed via Cronbach's (1972) generalizability theory.

Many other researchers have tried to collect data on classroom time use and language use via interviews and self-reports. For the proposed new studies, we reject this approach because of the lack of consonance between the teacher's intent (sometimes quite idealistic) and what he or she is actually able to accomplish. Dr. Flora Rodriguez-Brown, in conversation with the author, cited the anecdote of an Anglo teacher in bilingual education who claimed in interviews that she taught almost entirely in Spanish, but under observation, turned out to begin Spanish in the morning but to revert to English as she became fatigued. Her morning class, therefore, received a significantly different modality instruction than her afternoon class, despite good intentions. The paper data instrument, however, reported equal morning and afternoon treatments.

The teacher in this case (and in most others) was not dishonest, but rather, simply, very human. Certainly no blame can be attached to her, but the anecdote serves to illustrate the fact that in a large policy research study, it would be unwise to put too much credence in self-reports of

classroom process. Data collection of this type should be collected but used only in conjunction with more direct assessment.

Examination of In-School Materials

Many materials already available in the school will be useful for analyses. These include curriculum materials, classroom equipment and materials, teacher work plans, teacher evaluation materials, etc. While no study could be based solely on such information, it is valuable support for that gained by more obtrusive means.

These materials should be thought of not as measurements of process, but as a part of the process itself. They can be used directly in analyses. For instance, teacher plans for daily time use can help to establish the ideal with which the observed real can be compared. Discrepancies can be detected and analyzed by interview methods. Some of the specific real world barriers to implementation will be identified and illuminated.

Table 4

Classroom Level Educational Process Variables
Used in the AIR Study

Process Variables
Average Teacher Hours per Student per Week
Average Adult Hours per Student per Week
English as a Second Language Is Taught
Spanish as a Second Language
Whole-Class Instruction in English as a Percentage of Total English Instructional Time
Individualized Instruction in English as a Percentage of Total English Instructional Time
Group Instruction in English as a Percentage of Total English Instructional Time
Average English Hours per Student per Week
Grouping Index for English
ESL As Percentage of Total English Time
Spanish Use Allowed in English Classes
Whole-Class Instruction in Spanish as a Percentage of Total Spanish Instructional Time
Individualized Instruction in Spanish as a Percentage of Total Spanish Instructional Time
Group Instruction in Spanish as a Percentage of Total Spanish Instruction Time

(continued)

Table 4 (cont.)

Classroom Level Educational Process Variables
Used in the AIR Study

Process Variables
Average Spanish Hours per Student per Week
Grouping Index for Spanish
SSL as Percentage of Total Spanish Time
English Use Allowed in Spanish Classes
Whole-Class Instruction in Mathematics as a Percentage of Total Mathematics Instructional Time
Individualized Instruction in Mathematics as a Percentage of Total Mathematics Instructional Time
Group Instruction in Mathematics as a Percentage of Total Mathematics Instructional Time
Average Mathematics Hours per Student per Week
Grouping Index for Mathematics
Spanish Use Allowed in Mathematics Class

Another direct use of such materials (along with interviews and questionnaires) is to help to establish the context within which ethnographic observations can be planned. The observer and the teacher should also become acquainted to establish rapport as a social unit. All qualitative researchers appear to agree that ethnographic observation should not be done without such preliminary efforts. The trained ethnographer will not rely on such information, however, but will confirm it by his own observation.

Direct Observation

Direct observation methods are an essential component of new bilingual education instructional features studies, if these studies are actually to produce valid new knowledge. There are several domains that can be addressed only through observation techniques. These include linguistic use, participation structures, counts of significant events, and classroom interaction. Classroom time use can be assessed through indirect methods but, we argue, not very reliably.

Laosa (1977) stated that the process of bilingual education schooling can be best assessed by direct observation of what actually occurs in the classroom, particularly with regard to studies concerned with the degree of educational opportunity that ethnic minority students experience. He cites a few observational studies (Byers and Byers, 1972; Jackson and Laosa, 1974; U.S. Commission on Civil Rights, 1973; Rubovits and Maehr, 1973) that indicate that teachers behaved less favorably to ethnic minorities and that majority children's behavior towards majority teachers was more likely to elicit desired responses from teachers (since the children had a better understanding of implicit patterns of expectations and meanings of behavior). These studies also show differential behaviors by SES, academic achievement, and speech characteristics. Laosa's own (1977) study left no doubt that ethnicity and language had a significant effect on classroom interaction.

Thus the question as to whether or not there should be direct observation appears beyond argument. The question is, rather, what the methodologies should be, and how they should be integrated in a comprehensive system. There are at least three separate traditions in direct observation of classroom process that must be considered in designing the instructional features observation system. They are the interactional approach (e.g., Flanders, Medley, Stallings, etc.), the time-use approach (e.g., Cooley and Leinhardt, Wiley, Berliner et al.), and the ethnographic approach (e.g., Carrasco, Mehan, Cazden, Erickson, Rist).

Interactional and time-use approaches represent the quantitative, conservative tradition in classroom observation studies, although most of these studies are rather recent. The success of time-use approaches seems clear (i.e., there do exist relationships between the way instructional time is spent and children's test performance), but the usefulness of interactional approaches is less well-documented. Many studies have yielded either null effects or counter-intuitive patterns of significant effects (see Medley, 1977).

The following section, contributed by Gaea Leinhardt, reviews instrumentation issues for quantitative observation systems.

Quantitative Observation Systems (Contributed by Gaea Leinhardt)

Many different instruments and techniques are available for the measurement of the educational environment. Some are designed to capture the instructional event, and others the entire educational process. In this section, we review instruments in the latter category. Most were designed not as evaluative tools, but as descriptions of the interactive instructional process. The most complete listings of observational instrumentation are provided by Simon and Boyer, 1970, and Borich and Madden, 1977.)

Observation of classroom instruction can utilize techniques varying from running protocols, to frequency counts of teacher or student utterances to fine grained studies of body postures. It is common to distinguish high-inference codes, such as ratings, from low-inference codes, such as time and event sampling. A more substantial distinction can be made between instruments that focus on structural features, or on content elements of the environment. Instruments can also focus on single environmental elements such as the teacher, student, or curriculum or on the entire instructional domain, that is, all the elements and their interactions.

Structural or content-free instruments use the same constructs or measures across instructional content areas. That is, the same structural features are followed in all situations. The advantage of such an approach is causal parsimony. If a single set of instructional features can be found to influence learning regardless of subject matter or level, then great strides will have been made towards understanding the instructional environment. This would be invaluable in the bilingual instructional features study.

Structural Instruments

The most common type of observation tracks the verbal behaviors of teachers and (occasionally) students in classrooms. This approach grows out of the conviction that the main stuff of the classroom is what goes on between teachers and students or occasionally between student and student. This type of instrumentation is structural and focuses on elements in the environment, rather than the total instructional setting. Such instrumentation is often embedded in a package of larger instrumentation that is used to examine multiple elements (Cooley and Leinhardt, 1975b; SRI, 1972; Sustaining Effects Study, 1976; Marliave et al., 1977). Adaptations of these techniques could be useful in the instructional features study.

Work on teacher/student interactions has a clear-cut genealogy starting with the sociological work of Bales (1951) and continuing through Flanders (1966, 1967), and Medley and Mitzel (1963). In the post-World War II era, sociologists, including Bales, showed increasing concern for the styles of interaction processes. This interest stemmed from a desire to understand how a group could influence the behaviors and beliefs of its members in constructive and destructive ways. They posited that one way to determine this was to examine group interactions to see both who was talking and what the function of the talk was. The goal of this examination was, in part, to characterize the group as democratic, authoritarian, directed, etc. In adapting this approach to the classroom, Flanders (1966) retained the flavor and, in some cases, the original definitions of Bales' approach.

Interaction Analysis in the Classroom (Flanders, 1966, 1967) became the basis for the majority of classroom observation studies. This system also became the basis of much of the teacher effectiveness research in which little or no attention was paid to the rest of the instructional environment. The original purpose of Bales' instrumentation had somehow been lost. As many have documented, the results of this work were not encouraging; however, excellent reviews of the specific variables and their relationships to achievement do exist (Medley, 1977; Rosenshine and Furst, 1978).

A second class of instrumentation attempts to capture most of the instructional environment. Instructional instrumentation differs from teacher- and student-based instrumentation because it includes information about the total instructional environment including verbal interactions. For example, it may include information about assignment practices, classroom management procedures, frequency of testing, and grouping patterns. Instructional observation usually requires multiple instrumentation, such as direct observation, interviews, and logs (Leinhardt, 1976). The basic view is that things that influence instruction are more than the teacher/student interaction, but less than everything that's going on in the environment.

There are two compatible ways in which such instrumentation is developed. In one, observers are sent to the field to take protocols of the entire environment. These protocols then serve as the basis for coding or quantifying the environment. Examples of the elements that emerge are abruptness, attending, defiance, engagement, and oneness (Tikunoff, Berliner and Rist, 1975, p. xii). A second approach starts with a model of instructional processes and uses site-based visits to develop instruments to fit model and local conditions (Leinhardt, 1976; Cooley and Leinhardt, 1976 a). Examples of elements of interest are peer tutoring, speed of feedback, cognitive contacts, management contacts, and frequency of testing.

Structural instrumentation reflects the perspective that those dimensions of instruction that are facilitative or detrimental to academic growth are the same regardless of the context. The outcome of such instrumentation has tended to yield confusing results in terms of inconsistency across subject matter or grade levels. For example, in the Instructional Dimensions Study (IDS) (Cooley and Leinhardt, 1978), the variable that reflects the presence and use of a strategy to match students to their curricula correlates with student gain in the following way: -.03 in first grade reading, .25 in first grade math, .02 in third grade reading; .26 in third grade math (Cooley and Leinhardt, 1978). We would expect such inconsistencies to appear in studies of bilingual education instructional features.

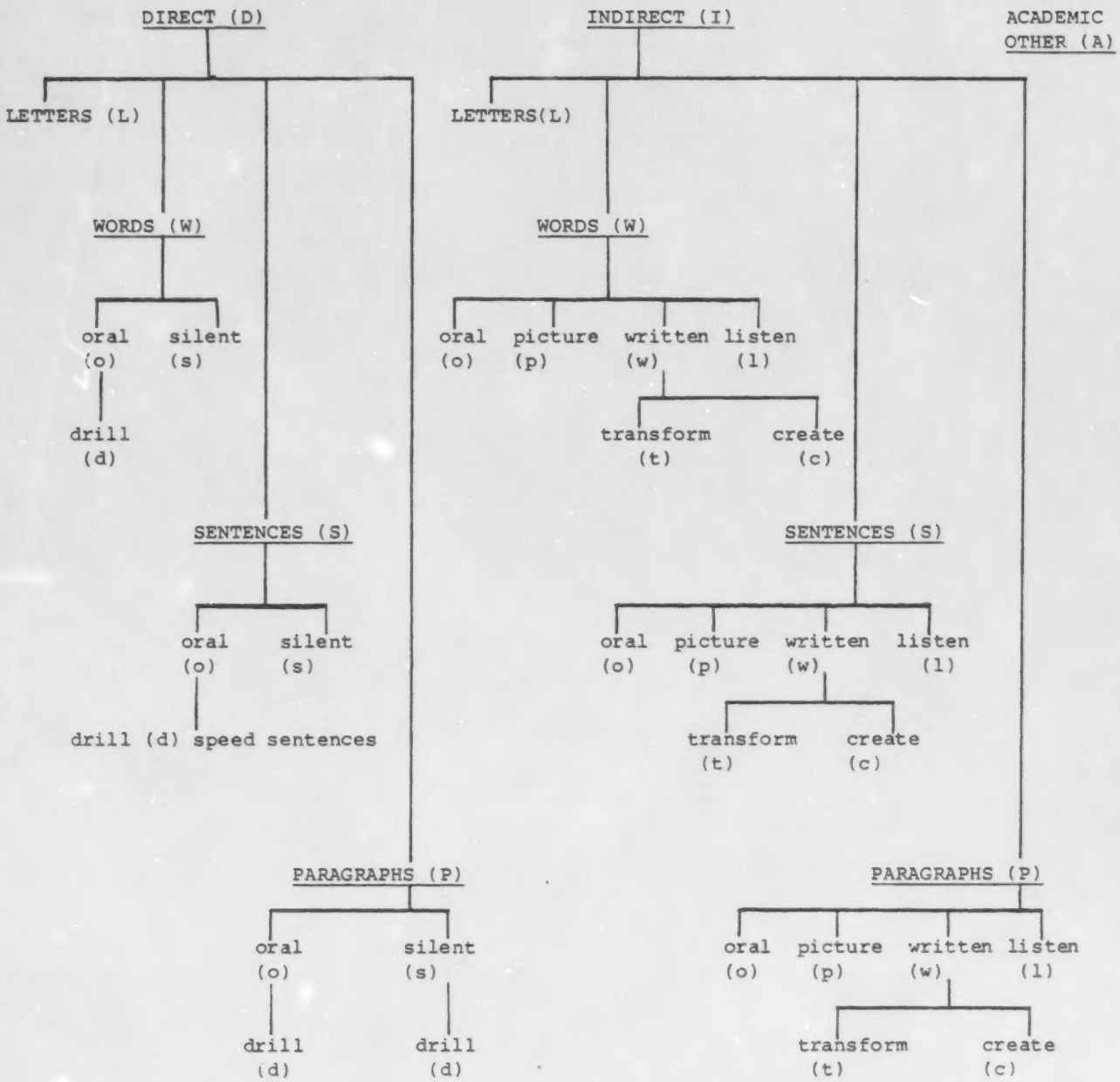
Content-Based Instruments

Instrumentation that reflects subject matter and grade level has recently begun to emerge. The first large-scale effort to incorporate content information grew out of the Beginning Teacher Evaluation Study (BTES). The model requires observation of academic learning time (ALT), which depends on the subject matter studied. ALT is different from percent on task because the observer, not the teacher, determines the tasks to

include. The BTES study defined "on what task" in terms of commonly used subject matter divisions such as long vowels, decoding, oral reading, silent reading, reading below test level, addition and subtraction without regrouping, with regrouping, etc. While these labels are neither mutually exclusive nor analytically based, they represent the first major attempt to examine what children are actually doing.

A second content-based approach that has been pursued can be called "analytic." This approach was developed in a study that focused on reading instruction in learning disabilities classrooms (Cooley, Leinhardt and Zigmond, in preparation). In this study, the reading behaviors of 60 students were followed for 30 to 60 periods over ten weeks. In learning disabilities classrooms, students study "reading" for 3 to 5 periods a day. Obviously, all that goes on in these periods is not reading. It became important to distinguish between reading, reading-related, and unrelated behaviors. Figure 2 presents the diagram around which observational categories were constructed.

Figure 2 is read from the top down. Across the top are a set of exhaustive classes of student behaviors for a student that is present. Reading behaviors are classified as direct (actual reading) and indirect (reading-related). Direct reading is further broken down into letters, words, sentences, and paragraphs, and oral vs. silent reading. Indirect reading is also broken down by level and type of activity, language processing (oral, listening) or symbolic manipulation (written). The observation system was designed to capture the stimulus presented to the child, the responses the child makes, and the setting in which that activity occurs. This system yields information on how programs actually teach reading, not just how much time a student spends on task during reading period. When this type of observation schedule is combined with other instrumentation, one has a content-based and instructional environment estimate of treatment.



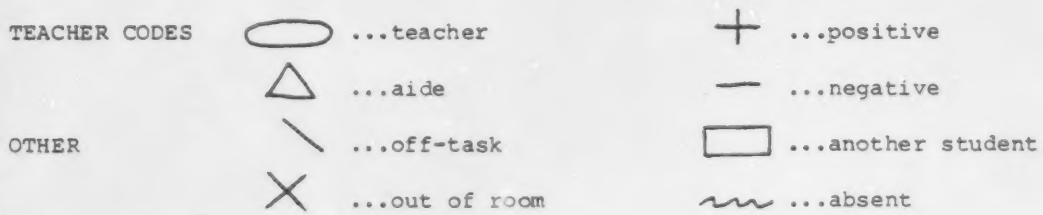


Figure 2 Analysis of reading behaviors(Leinhardt)

The limitations of the content-based approach are that one level of generalizability is lost, and that findings are restricted to grade level, subject matter area, and the underlying learning model of instruction that is used. The BTES employs content-based instrumentation for students but includes structural components to measure other parts of the environment.

It is useful to review categories of activities and actors and in relation to types of measures. Table 5 displays six potential foci: community, school, curriculum, classroom, teacher behaviors, and student behaviors. For each focus, a small sample of the types of measures implied for that target and the models that suggest the target are listed. While all of the measures listed could be obtained by direct observation, the first four foci can probably be measured with sufficient accuracy by interviews or questionnaires. Teacher and student behaviors, however, require direct observation.

Focus	Sample content Areas	Models
Community	Home Environment	Wiley & Harnischfeger
	SES	
	Community Attitudes towards School	
	Attitudes towards assimilation/ maintenance	
	Language Used	
	Aspirations and Values	
School	Environment	Wiley & Harnischfeger
	Size of student body	
	Special programs	
	Demographic descriptors	
	Stability	
	Leadership styles	
	Relation of bilingual education to the larger school environment	
Curriculum	Nature of the task	Bloom
	Overlap with criterion	Cooley & Leinhardt
	Organization	
	Level of difficulty	Bloom, Cooley & Leinhardt
	Pacing	BTES
	Language Used by Content Area	
	Language Varieties	
Classroom	Organization	Wiley & Harnischfeger
	Attendence	Cooley & Leinhardt
	Press	
Teacher Behaviors	Content of behaviors	All models
	Rate of behaviors	
	Planning	
	Instruction	
	Feedback	
Student Behaviors	Content of task performance	All models
	On task	
	Entering behaviors	
	Exiting behaviors	
	Quantity of exposure	

Table 5. Areas for measurement of treatment

Reliability

Clearly, if one observes the instructional environment to examine the dimensions along which instruction varies, one must be seriously concerned with the question of the reliability of what one is recording, although the evaluation of educational treatment must not turn into an exercise in validating instrumentation. (cf. Medley & Mitzel, 1958, 1963; and McGaw, Wardrop, & Bunda, 1972, with the suggestions of Rowley, 1976.

Before considering specific types of reliablity, the question of "reliability of what?" should be examined. Rowley (1976) makes several useful distinctions. In order to collect observational data, an instrument is used. An instrument helps to generate an observational record, i.e. a sheet with recorded codes. From these records, measures are constructed, for use in the actual study. Measures generally incorporate one or more codes that appear on a record; thus, codes can be grouped to form measures that have considerably higher reliability than do the initial codes . For training procedures in observational studies, it is important to get high agreement at the code level. But for estimating reliabilities for data analytic purposes, it is appropriate to attain reliability at the measure level.

Reliability at the code level encompasses two main issues: the agreement between observers, and the generalizability or stability of the characteristics being observed. (Medley and Mitzel, 1963; Cronbach, Rajaratnam, & Gleser, 1963; McGaw et al., 1972; Frick and Semmel, 1978; Erlich and Shavelson, 1976; Rowley, 1976, 1978).

Inter-observer reliability, the degree to which two (or more) observers observing the same event obtain the same measures, is well within the grasp of current technology. It requires an instrument and manual that has

well-defined codes and clear definitions of when and when not to code an event. Obtaining inter-observer reliability also requires that a systematic training procedure be used, one which exposes observers to realistic obsevational situations. Finally, it requires that reliability checks be continuous throughout the course of the study to avoid delay in training (Reid, 1070) and that observers be rotated across settings.

The most comprehensive approach to these issues is that of generalizability theory, developed by Cronbach et al. (1963). This theory can be viewed as an extension of classical test theory, in which only two sources of variance are recognized (true score and measurement error), to the case of multiple sources of variance. These sources include the observer, instability in the process itself, subject-to-subject and class-to-class variation, and other possibilities. The statistical technique involved is that of variance components analysis.

The great limitation of the method is that, if estimation procedures are not to be extremely complicated, it is necessary to set up a preliminary ANOVA design data collection effort (G-Study) on which to base estimation of variance components. Goodrich and Singer (1978) have developed some new techniques that use less structured data, but there are certain methodological disadvantages. In any case, generalizability analyses can be used to determine the effects of observer disagreement and length of observation on constructed variables, and thus to provide a basis for planning the observation schedules that achieve required reliability at minimum expense.

The BTES study used these procedures at an essentially state-of-the-art level, both formatively and summatively. This example should be followed in designing quantitative observation systems for bilingual education instructional features studies.

Concern with issues of reliability at the code level may lead the researcher to design a sampling pattern that is more likely to capture the phenomena of interest. Rowley (1978) suggests that one should increase the number of separate observations rather than merely increasing the duration of single observations. Thus five 10-minute observations on separate days is superior (in terms of reliability) to one 50-minute observation. Rowley attributes this to an increase in independence of observational measures. He refers to "observational links" which are more likely to be present if one observes on the same day, for example, than if one observes on separate days. However, a more useful distinction is to consider the variety of information obtained by different samplings, or the unit to which one wishes to generalize. One could observe five 10-minute periods consecutively (one 50-minute period), or one could observe the first 10 minutes of every hour for five hours during the day. Presumably, the first 10 minutes of every hour yields more information about hours and days than a single 50-minute period, but it's also presumable that 10 minutes on separate days will yield still more information. There is a trade-off between the length and frequency of observation. In general, it is better to opt for more frequent observations than longer ones.

Issues of validity are closely related to issues of reliability. Obviously, it is important that the instruments used in a study accurately assess the salient features under scrutiny. No instrument is of value if it fails to measure the right traits. A complex method for determining validity (multi-trait multi-method), has been developed, that goes beyond simple face validity (Campbell & Fiske 1959; Borich, Malitz, Kugle, & Pascone, 1977). One examines a specific element (trait) and asks whether values in that element are more similar to the same element measured a different way or to different elements measured the same way.

Ethnographic Approaches

As discussed in Chapter 3, purely quantitative observation techniques will not suffice to provide information about classroom process. The need to examine participant structure, differential interaction patterns, and patterns of linguistic use demands an ethnographic approach. The chief technical problem in designing an observation system is to combine quantitative and qualitative approaches. This section provides an examination of some aspects of the qualitative side.

Note that the role of theory and specific hypotheses in ethnographic interpretations of observations is mildly controversial. While some feel that the method is hypothesis-free, others (e.g. Mehan, private conversation), feel that the hypotheses are always there, at least implicitly. Mehan further feels that the theoretical framework should be explicit and "up front," but that development of the system should be "bottom-up."

There are two basic methodologies that can be employed in instrumenting an observation system - - we can rely on in-class observers who record data in code or prose on the spot, or we can rely on in-class videotape and audiotape records for later coding and analysis. Each method has strengths and weaknesses, and affords a wide variety of alternatives. Hybrid systems are also possible and may be a reasonable choice.

The observation system must in any case contain traditional interactional, time-use, and ethnographic components. Interactional and time use components would follow classical coding schemes. There is a wide variety of possible ethnographic coding systems ranging from simple English prose to numerically coded global (macro) judgments. Ethnographic systems would also include interactional content.

The following are factors in making a decision between the two types of systems. See Mehan (1978), Rodriguez-Brown (1976).

Simultaneity

For full integration of interactive (event oriented), time use, structural and ethnographic analyses, observations should be simultaneous. Otherwise relationships between the three analytic tracks would be attenuated by natural temporal fluctuations in the processes, and by varying context. Multiple in-class coders would be required and their records synchronized, so that track can act as a context for the others. The inclass coder system is therefore clumsy, intrusive, and probably infeasible. Of course, in-class observers should in any case do preliminary ethnographic analyses to help ground them and establish a theoretical causal model.

Videotape, on the other hand, promises automatic simultaneity. As in Carrasco's approach, the tape can first be analyzed for context, then for macroethnographic content, interaction events, and time use, referred to the context. Of course, the entire tape must also be referred to time of day, the curricuulum plan, organization, etc., to establish the global context.

Reanalyzability

Tapes may be analyzed and reanalyzed in a variety of ways (e.g. ethnographic, microethnographic, time use, language use by subject area, interaction analysis, "rare events" records, investigation of differential behavior (including segregation). Analyses can be as deep or shallow as deemed appropriate. Any significant instant can be explored in great detail. Observer reliability is thus not the issue that it is in in-classroom coding system (e.g. that of Stallings). Coder reliability can be conveniently assessed and quality measures made. The work of less able coders can be redone. Full scale generalizability studies (Medley and Mitzell 1963); Cronbach et al., 1972 can be conveniently performed at leasure. These can address not only coder reliability but the stability of phenomena under observation.

In the National Day Care Study it was determined that in-class coder observations were substantially influenced by the characteristics of the observer, especially by his or her race. The ethnicity of the coder in a bilingual education study would almost certainly have similar effects. Videotape would allow minimization of observer characteristics as a confounding factor, by appropriate assignment of coders.

Language Samples

Language samples can be collected via wireless transmitters attached to the children. If video tape is used these samples can be keyed electronically to the three track analysis to determine the precise context at each instant.

Fidelity

Although videotape observations must be made by persons trained in videotaping techniques as well as ethnography, they will still be somewhat influenced by point of view in the classroom. However this situation is not nearly so severe as for in-class coding methods. The videotape record is on the whole likely to be a more faithful primary data record than the in-class coded record, although it may not cover the entire classroom.

Expense

Videotape equipment and materials are more expensive than those required for in-class coded records. Camera operators must be trained both in taping methodology and ethnography, and must understand both English and the native language. Videotape equipment is bulky and may elicit "reactive" classroom behaviors On the other hand in-class observers also induce these behaviors, although perhaps not to the same degree.

Volume of Materials

The volume of ethnographic records in a large scale study can be overwhelming. Special information storage and retrieval methods must be devised to analyze the data feasibly. To use videotape in the proposed new study would require state of the art advances in this area, particularly in computer based retrieval systems. This problem is also present for in-class coders but is probably less severe.

Permissions Problem

It may be necessary to collect permissions in the classroom on a child-by-child basis. If one or more children refuse then they must either be sent home on videotape days or sent to alternative classrooms. Either "solution" might produce further family or community reactions. Thus the classroom itself may be effectively lost to the sample. If many such classrooms exist then the validity of the sample would be seriously compromised. If in-class coders are used they may simply skip children for whom permision has not been granted.

4.2 Measures of Language Proficiency

There are several methods (see, e.g., Sanchez et al, 1978) by which language proficiency can be assessed, of greater or lesser validity and reliability. These are:

Tests
Interview-based direct rating
Video-tape language samples
Self-report
Teacher-rating

These measures can vary in complexity from simple unidimensional scales (i.e. a vocabulary count) to complex multidimensional psycholinguistic scales. Moreover they can be single domain (e.g. the classroom only) or multiple domain (e.g. the classroom, community, and home).

An argument for multidimensional scales was provided to us in personal correspondence by Christina Bratt Paulston. She points out that there is no overall "G effect" in language learning. Pronunciation and communication skills do not correlate with IQ, while reading, writing, and vocabulary do. Therefore, measures that do not break apart those two areas are missing something fundamental in language learning.

Measures of English Proficiency (Contributed by E. Cervenka)

Instrument Development

A severe constraint on conducting large scale research studies of children in bilingual education programs in Grades K-6 has been the general inadequacy of language and language-related assessment instruments. A comprehensive review of currently available instruments finds that a great number of tests and other measures relating to bilingual education have been developed since the passage of the Bilingual Education Act of 1968; however, their adequacy in terms of, for example, APA standards remains in serious question. Lock, Pletcher and Reynolds noted in Assessment Instruments for Limited English Speaking Students: A Needs Analyses (1978):

> "Concern over the adequacy of available instruments to assess the academic progress of limited English-speaking students has been voiced in almost every major study or conference on bilingual education during the last several years."

In its 1975, 1976, and 1977 Annual Reports to the President and to Congress, the National Advisory Council on Bilingual Education recommended the development of adequate testing methods and instruments for limited English-speaking

children. The General Accounting Office's report to Congress, Bilingual Education: An Unmet Need (1976) recommended an examination of "the appropriateness of academic testing instruments available for children with limited English-speaking ability" and, if needed, that action be taken "to have better ones developed". The U.S. Commission on Civil Rights report, A Better Chance to Learn: Bilingual Education (1975) found that bilingual education program implementation was hampered by a scarcity of adequate assessment.

However, the overall state of the art of language test development for children K-6 should not come as a surprise if one considers the following:

1. Until a decade or so ago, language was not really considered a factor in education or education research because English was, or was considered to be, the language of the school; thus, language was taken for granted.
2. The flurry of activity in the area of test development for bilingual education undertaken mainly by individuals is only about a decade old.
3. The development of adequate language tests, from item selection to demonstration of reliability and validity, require multidisciplinary skills; e.g., those of linguistics, psychology, statistics, etc. Unfortunately, these skills are not involved in many instances of test development. Currently, most language tests suffer from inadequate application of psychometric expertise; e.g., item analysis, field testing, reliability studies, validity studies, validity studies and provisions for interpretation of scores.
4. Adequate language test development is not only a matter of a long term effort, it is also a matter of great cost. Troike (1978) estimated that only .5 percent of the total ESEA Title VII budget since 1968 has been devoted to research. Probably much less has been spent on test development. The Office of Educational Evaluation of the New York City Board of Education estimated that $500,000 was spent on developing the Language Assessment Battery during the first year only.

It is conservatively estimated that the Bilingual Testing and Measurement Component of the Boston Public School spent over $400,000 in developing the Placement Evaluation Battery over a three-year period.

5. Finally, development of language tests for children is difficult because children - particularly young children - are often less cooperative and responsive than adults in test situations and their psycholinguistic development is only partially known.

There are at least a dozen publications which give descriptive, uncritical reviews of assessment instruments for bilingual education. Only two publications focus on critical review of instruments: Oral Language Tests for Bilingual Students: An Evaluation of Language Dominance and Proficiency Instruments (Silverman et al., 1976) and A Guide to Assessment Instruments for Limited English Speaking Students (Pletcher et al., 1978), but the critical reviews of these publications leave a great deal to be desired in terms of completeness and user information. Fortunately the Mental Measurements Yearbook, 8th ed., included a number of language test reviews which were of good quality and utility.

In their needs analysis of assessment instruments, Lock, Pletcher and Reynolds (1978) argued that two priorities be established for test development:

1. Because educators often appear ambivalent toward publishers' attempts to develop unbiased instruments, and because many bilingual educators are unaware of potentially useful assessment instruments, a top priority for assessment instrument development is the improvement in communication between instrument developers and instrument users. (p.3).
2. Of only slight lower priority is the need for the refinement of available assessment instruments. (p.4)

Probably the most important study problem presented by research on multiple languages hinges on the existence of valid and reliable language proficiency and academic performance measures for each culture. Where no

adequate instrumentation is available, it will need to be produced by refinement of existing instrumentation, translation or development of new instruments. Approaches can be employed that do not rely heavily on quantitative assessment. It may also be wise to separate language cultural groups that cannot support generalizable analyses from those that can.

Test Administration for Bilingual - Bicultural Linguistic Minority Children

This subsection deals with some of the characteristic problems of testing language minority children. The usefulness of a test given to bilingual - bicultural children and its resulting data depend in large part on appropriate and intelligent administration.

The central set of problems relating to the administration of assessment measures, whether standardized paper and pencil tests or observational measures, can best be grasped by viewing a "test situation" as a communicative event. Some two decades of research on how human beings communicate and interact have led to the conclusion that substantially more is involved than that which is mediated by the structure of language. Hymes (1964) has labled this large study of communicative events, "the ethnography of communication," and situating within this the study of speech events, "the ethnography of speaking". Hymes (1964) went on to note that "the starting point is the ethnographic analysis of communicative habits of a community in their totality, determining what count as communicative events, and as their components, and conceiving no communicative behavior as independent of the set framed by some setting or implicit question." Within the field of the ethnography of communication, a number of subsystems of communication have clearly been delineated, besides the language. These subsystems include paralanguage (roughly "tone of voice phenomena"), kinesics (body motion communication, gestures, posture, stance, manner, etc.) proxemics (cultural use of space and distance as a signalling system, oculesis (eye contact, eye avoidance in social encounters, etc.) and a number of other such culturally-determined communication subsystems.

The communication subsystem encompassed by the "ethnography of communication" present little or no problem when the test administrator or observer and the individuals being assessed share the same overall system of communication. Test administration problems arise when test administrator and the examinees do not share this overall system of communication or when the administrator is not aware of it and does nothing to compensate for a partial breakdown in communication with the consequences that unintended variations are introduced into test scores and measures.

In order to minimize such variations in scores, several steps seem to be called for. First, it will be necessary to conduct small scale ethnographic studies of test-taking and test-giving within the various language and cultural groups selected for study. These studies will be useful in standardizing testing administration setting and training test administrators to give tests in approximately the same way. Secondly, it will be highly desired that all test administrators be native or near-native speakers of the languages they use in test situations. Probably for some portion of test administration involving spoken stimuli, tape presentations may be utilized.

It is to be reiterated that the point of taking into consideration the other subsystem involved in communication in the test administration situation is not to be "fair" to children, but to minimize unintended variations in scores and thus to improve measurement quality.

The SWRL Linguistic Skills Framework

One of the most important recent developments in BE research is the preparation of a Linguistic Skills Framework by Southwest Regional Laboratory under contract with USOE. The framework specifies the linguistic skills necessary to function effectively in an all English speaking classroom, grade by grade from K - 6. The framework is hierarchical, i.e., it classifies specific skills under three domains. Reading, writing, oral, and aural areas are all considered. At this time SWRL has completed the

specification of skills and has provided sample test items, but no test per se exists. SWRL is Planning to use the LSF in preparing a placement system for individual Title VII projects. Actual tests based on the LSF are years in the future and cannot be considered for BE instructional features studies being planned currently.

The SWRL work-in-progress does not actually break new ground, nor are the materials ready to use. But in the opinion of at least some linguists (e.g. Edward Cervenka), the SWRL work is the best, most comprehensive specification for a test of English Language proficiency. Therefore we should not allow the lack of genuine innovation in the LSF to deter us from bringing this important new work to bear on instructional features studies.

Because of the importance of English language proficiency testing in nearly any bilingual education study, and because of the quality and scholarship of SWRL's work, the option of delaying English language testing until a SWRL-based test is available is attractive. Even if off-the-shelf measures were used, the first conceivable Fall testing is in Fall 1981. There is no possibility that large scale testing of any kind could be arranged for Fall 1980. An RFP for the study could not be let before approximately June 1980, and this would not allow sufficient time for making the necessary arrangements through FEDAC, CEIS or with local LEA's. Thus, waiting for a SWRL-based instrument would delay data collection for at most one year. It is probably worth while to allow this delay for the sake of producing a valid test of English Language Proficiency.

Determination of Language Dominance

Determination of language dominance has been an issue in both research and local evaluation studies, and is a factor in designing the instructional features study. One issue is whether a dominance determination of children in the study should be made in addition to assessment of proficiencies in both languages. It is clear that dominance adds an independent behavioral dimension to the picture but not that it is a dimension essential to the study.

If language dominance is deemed necessary, then the measurement problem must be faced. Techniques have involved (1) teacher rating, (2) word association, (3) flexibility tasks, and (4) reaction time. Teacher ratings should be excluded from consideration because the child's in-school use of language is only the "tip of the iceberg" (Zirkel, 1974) of the child's total usage. It may be possible to construct an approximation to dominance via proficiency scores in two languages. Zirkel, for instance, has proposed the use of parallel aural ability tests in the two languages: the difference of the two scores becomes a measure of dominance that, however, may ignore the child's linguistic behavior.

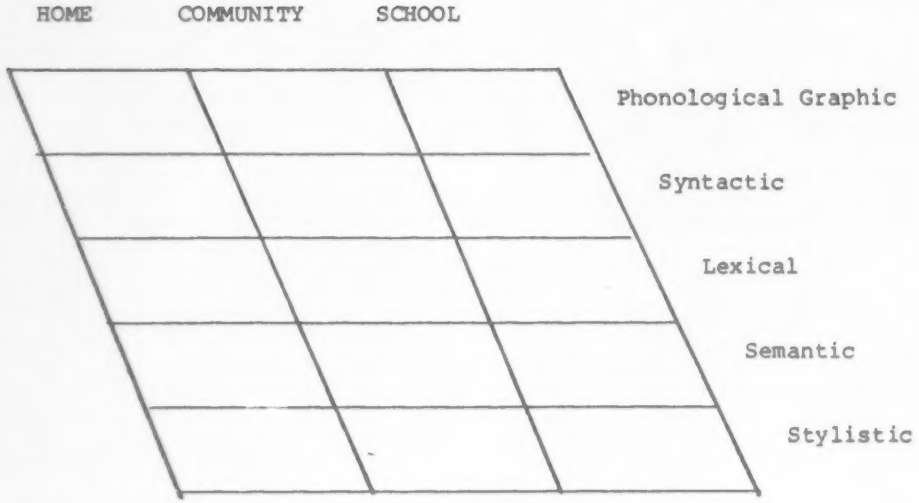
Figure 3 shows a graphical representation of Spanish/English dominance from Zirkel (1974). It emphasizes the multi-dimensional aspects of the problem in linguistic domain, context, and modality (listening, speaking, reading, writing.) In planning the instructional features study it will be necessary to specify which domains must be assessed.

Individual Monitoring of Discourse

An important source of data is the stream of discourse by individual children and teachers. This can be used for interaction analyses, analyses of extent of participation, and for samples (codable). It is sometimes difficult to pick up individual discourse via videotape because of background noise and the confusion of multiple voices. An ideal way to get the data appears to be that of "wiring the children" with radio transmitters strapped to the back. It might seem that this method would be resisted as invasive. Two sources (Carrasco, Dominguez, SEDL) have informed us, however, that the "children love it" and vie to be the lucky subjects. The parents are the potential source of objections, but these are easily overcome by in-home or in-school demonstrations where it is made clear that the procedure is non-threatening and fun for the children (Dominguez, SEDL).

REPRESENTATION OF BILINGUAL DOMINANCE MATRIX

from Zirkel (1974)



LISTENING _____

SPEAKING _____

READING _____

WRITING _____

(CULTURE) _ _ _ _ _ _ _ _ _

SPANISH <---- ENGLISH ---->

Figure 3

Chapter 5

MANAGEMENT, ORGANIZATION AND PHASING

Previous chapters of this report dealt with technical and substantive factors that should be considered in planning the Instructional Dimensions Study. The topics of this chapter are equally, or even more, important to the success of the study, but they are harder to discuss at this stage of planning. We cannot provide meaningful specifics on study organization until we know the specific components of the study. Our discussion in this chapter is therefore very general. The issues involved will be dealt with specifically in later reports.

5.1 Organization of Study Components

The overall design for BE instructional features research - if it is in fact to break new ground and improve substantially on previous work - will be complex. It will require skills from a number of disciplines, including ethnography, anthropology, statistics, data management, quantitative analysis, instructional dimensions research, sociolinguistics, and psycholinguistics. It will require personnel from a number of ethnic and language backgrounds. Almost certainly, there will be several different but (not disjoint) studies and substudies that require different skills and qualifications of the contractors. Although these components may be modular, they must all be linked into an overall pattern. This obviously raises special problems for study management.

Some study components that might be included are the following. These are not intended as actual design alternatives but as an example of the organizational problems that might arise.

(1) A large scale quantitative/qualitative study of in-classroom process and features.

(2) Smaller scale ethnographic/anthropological substudies designed to answer specific questions, or to explore particular domains that do not lend themselves to large scale study.

(3) Supporting focused substudies in particular areas.

(4) A consumer study of attitudes and objectives to be based on interview and questionaire data from children, parents, teachers, community people, congressmen, and others.

(5) A contextual features study of the support systems afforded the bilingual child outside the school, i.e., by neighborhood institutions, churches, ethnic clubs, and family.

(6) A longer term follow up study of students' later success academically, economically, and socially.

If NIE issues a single RFP, it would attract proposals in which most respondents would bid numerous consultants and one or more subcontractors for specialized components of the overall study. If several RFP's are issued there is a number of ways in which the overall study can be organized. In either case the study could involve individual consultants, consultant panels, university departments and ethnic studies centers, large profit-making and not-for-profit consultant firms, and specialized smaller firms. The next section deals with some of the issues in assembling these actors into a coherent organization.

5.2 Management

The management problem - one that NIE must resolve in Spring 1980 - is to determine an organizational plan for contracts, responsibilities, study components, and study phasing. Some models of how this complex job can be performed are presented below, along with commentaries on their strengths and weaknesses.

It is only fair to point out that these organizational approaches are seen from the point of view of a large profit making research firm(AAI). Of course, AAI has worked for the federal government under a wide variety of contractual arrangements and interdependencies with other contractors, consultant panels,and has considerable experience with the different problems that arise. The author of this report has discussed the pros and cons of different approaches with several individuals at AAI, and accordingly, a great deal of accumulated experience is reflected. NIE will, of course, make its final decision based on its own experience and perspective.

Model 1 NIE (Title I Evaluation Model)

In this model NIE designates an internal staff to manage and integrate a number of parallel contracts, with different contractors, large and small.

This was the model for the Title I evaluation conducted by NIE for the 1977 Elementary and Secondary Education ACT reappropriation hearings. This was felt to be a highly successful effort and was enthusiastically received. According to Zweig (1979), the following features were present:

(1) The policy-relevant questions were posed at the outset of the study in the legislation.

(2) The study was designed as a multiple-component, integrated enterprise with the lead agency, the National Institute of Education, the coordinator of a great many contractors of special analyses and evaluations.

(3) Continuing consultation with congressional staff was built into the study's schedule.

(4) Interim and final reports were geared to legislative events in the reauthorization cycle.

(5) Reports were written in language everyone could understand.

The direct control that NIE had over final study products was a great advantage, since NIE's beliefs and interests could be reflected directly in reports to Congress.

The principal disadvantage of this model occurs when some of the contracts are interdependent - for instance, data collection and analysis are by separate contractors. In this case the analytic contractor depends on the data collection contractor but has no direct authority over them. Communications must usually flow through the government project officers rather than in person-to-person contracts between contractor personnel. As a result, time delays and a general loss in efficiency sometimes occur. The data collection/analysis problem is discussed in detail below.

NIE has already implemented a plan of this type, at least to a limited degree. The three ethnographic sociolinguistic studies to be sponsored by NIE are examples of the type of research we term "supporting substudies" above. Their results should be integrated with those of the instructional features study, either by NIE (Model 1) or by a coordinating contractor (Model 2 or 3, below).

Model 2 ACYF (Social Competency Model)

In this model, NIE hires a management contractor, which could be a large research firm, a university, or a smaller firm with substantive capabilities. The management contractor is ultimately responsible for all products. They would have individual subcontracts with the other contractors. Alternatively, NIE might fund the other contractors directly but make lines of authority very clear. An example of this arrangement is the ACYF contract with Mediax to manage the Social Competency contracts.

The advantage of this model is that, since the managing subcontractor is legally responsible for products, they will not be hesitant to reorganize, cancel subcontracts as needed, discharge personnel, etc., in order to deliver the products they are legally bound to provide. A further advantage lies in the competitive nature of the RFP and contracting process. RFP's for the bilingual education instructional features studies will undoubtedly draw a number of bids, because of the large expected revenues involved and inherent interest in the work. The competition among large research firms has heightened in recent years, as markets have decreased in size, and technical professionalism and competency of contractors has increased. The instructional features RFP's will undoubtedly elicit some very strong proposals. NIE will be able to pick and choose the strongest.

A major difficulty is that NIE would loose immediacy of contact and up-to-the-minute awareness of what is going on since it would not be so directly involved. Moreover, it would sometimes have to rely on judgments of the contractor on matters that may be politically sensitive. There may also be modest additional costs for this model, since subcontracted studies bear double fees once at the subcontractor level and another at the prime contractor (where a subcontract is a direct cost item). The cost of management should not be very different than in Model 1.

Model 3 FNS (School Feeding Model)

In model 3, NIE lets a number of parallel contracts but appoints a single coordinating contractor. This is the similar to Model 2, except that there is no legal responsibility of one contractor to another. An example is the group of School Feeding Studies contracted for by the Food and Nutrition Service with Abt Associates Inc. (AAI), Systems Development Corporation (SDS), and Research Triangle Institute (RTI), where SDC has a special coordination role. The advantages and disadvantages are similar to those of Model 2, except that because the coordinating contractor lacks authority over the others, quality control must be NIE's responsibility.

One issue that has been a clear problem at times (but always one that was satisfactorily resolved) is dealt with in the following paragraphs.

In any of these models, data collection and data analysis can be functionally separated. In the bilingual education instructional features study, data collection techniques will undoubtedly include interviews, self-report questionaires, small scale ethnographic observation, larger scale quantitative/qualitative observation and some testing. Most of these efforts lie well within the capabilities of major contractors in the educational research market. But observation and testing on a large, national scale are specialized, difficult, technical tasks and should probably not be entrusted to a contractor without extensive previous experience. The few contractors with this experience do not necessarily offer the range of other skills and experience required for the overall job. Therefore NIE might consider a plan in which a large portion of the data collection - particularly observations and testing - is conducted by one contractor, while the data analysis and smaller scale observations are conducted by others.

There are two ways that arrangements could be made - through subcontracts or through parallel contracts with NIE (i.e. by models 1, 2, or 3). In cases where the separate pieces of the work are both large, it is more usual to let separate contracts. In this case, a large management burden may fall on the federal government. In either case, cooperation between the contractors must be assured by building liaison directly into the management structure.

It is Abt Associate's experience that difficulties sometimes arise in the parallel contract mode. In several cases AAI has been the data analysis contractor in conjuction with a separate data collection contractor. There was sometimes some mutual dissatisfaction, and the consequent need for negotiation through the project officer. In most cases AAI would have preferred an arrangement in which the data collection contractor was under subcontract to us, and hence subject to our management control.

Study Objectives

Of the five features listed by Zweig in the discusion of Model 1 above, only one (the second) is peculiar to Model 1. The other four must be present in any successful oirganization of study. NIE has already begun to address the third to fifth features, involving coordination with the needs of Congress, by dividing the overall effort into these phases: (1) this planning study, (2) implementation of the study itself, and (3) coordination and communication with Congress for the 1982 reappropriation hearings.

Some of the questions to be addressed by the study are posed in the legislation, but our conversations with NIE staff indicate that NIE's purposes are broader and extend to the area of basic research. Indeed, the field of bilingual education desperately needs such research. Of course, all of this research will ultimately bear on policy, but the balance of the short and long term goals must be resolved. This is particularly important in designing the phasing of this study, discussed in the next section.

5.3 Phasing

Major results of the study are to be made available by Spring 1982, in time for the 1982 ESEA reauthorization hearings. Only these results will have immediate policy impact. Conclusions made available later on would contribute substantially to the corpus of knowledge about bilingual education but would not be so useful in the short run. As we pointed out in Chapter 3, it will be difficult to conduct a high quality study in time for the ESEA hearings. A high quality study would involve instrument development, validation, and implementation, for perhaps a year; data collection over at least one year; and analysis that could extend from six months to a year. There is a clear cut trade off to be made by NIE between timeliness and quality.

In Chapter 3 we outlined some directions to data collection that stressed application of ethnographic techniques to large scale observational data collection. Taking this route will increase the length of time before major results are ready. It would not be possible to install working data collection, reduction, and retrieval systems for the 1980-81 school year. At a normal pace, RFP and contract/ grant award process will be complete by midsummer 1980. Development of new data and observation systems requires a tangible advance in the state of the act that is entirely feasible, but will require time. A reasonable plan for the high quality research would use the entire 1980-81 school year for development and installation of a data system and other important preparatory activities, including the following, which are to be considered only as examples.

(1) Development of a finalized sample design and selection of a sample.

(2) Site development work to secure cooperation with the studies to be conducted.

(3) Installation and shake down of an interview (or self report) information collection system.

(4) Performance of traditional anthropological/ethnographic studies in a subset of sites to ground development of the observation system, and to generate hypotheses.

(5) Development of a computer-based information storage and retrieval system designed around the observation system.

(6) Development and installation of a multicomponent observation system, to be ready by Fall 1981.

(7) Evaluation of a language proficiency assessment system, clearance through required authorities (e.g. CEIS), and securing of required LEA and community support, to be ready for data collection in Fall 1981.

Under this plan, the actual data collection would occur mostly in the 1981-82 school year. Reports for the use of congressional hearings could not possibly use data collected later than January 1, 1982, and that would allow only three months for analysis - a very short time in a complex educational study. If this route is taken NIE would have to identify a set

of interim results to be made available from early rounds of data collection, and to make then available in report form in March 1982. Then analysis would proceed but the results could not bear substantially on the 1982 reappropriation hearings.

We do not now know the degree to which measurement of change will be required in the instructional features study. Change measures might be used in analysis of the relation between consequences for children and instructional features, i.e. in a process/product paradigm. But if change measures are required there will be additional scheduling problems, which we discussed in Chapter 3 (The Time Frame and Measurement of Change). Meaningful analyses of this type would be different and perhaps impossible to make available for the 1982 hearings unless pretests were administered by Winter 1980-81.

Persons Contacted in Preparing this Draft

Tomas Arcieniega, School of Education, San Diego State University

Ray Chesterfield, Juarez and Associates, Los Angeles, CA

Sam Cohen, Babel Inc., Berkeley, CA

B. Roberto Cruz, Babel, Inc. Berkeley, CA

Malcolm Danoff, American Institutes for Research, Palo Alto, CA

Domingo Dominguez, SEDL, Austin, TX

Heidi Dulay, Bloomsbury West, San Francisco, CA

Ruben Espinosa, Institute for Cultural Pluralism, San Diego State University

Tracy Gray, Center for Linguistic Analyses, Arlington, VA

Carlos M. Haro, Chicano Studies Center, UCLA

Neal Killalea, Killalea Associates, Alexandria, VA

Jose Llanes, University of Eastern Michigan, Ypsilanti

Hugh Mehan, Laboratory for Comparative Human cognition, UCSD

Xenis Melendez, New York Board of Education

Carmen Mercado, Hunter College, New York

William Milan, Yeshiva University, New York

Luis Moll, Juarez and Associates, Los Angeles, Laboratory for Comparative Human Cognition, UCSD

Alberto Ochoa, Institute for Cultural Pluralism, San Diego State University

Masahito Okada, SWRL, Los Alamitos, CA

Raymond Padilla, Eastern Michigan University

Pedro Pedraza, Center for Puerto Rican Studie, New York

Ray Perez, Juarez and Associates, San Diego, CA

Donald Reisler, DBS Corporation, Arlington, VA

Flora V. Rodriguez-Brown, University of Illinois at Chicago

Marietta Saravia Shore, Columbia University

Rudolph Troike, National Clearinghouse for Bilingual Education, Arlington, VA

Jose Vasquez, Hunter College, New York

Bibliography

American Institutes for Research. Study of Bilingual-Bicultural Projects Involving Native Americans, Indo-European, Asian and Pacific Groups. Palo Alto, CA, 1975.

American Institutes for Research. The Identification and Description of Exemplary Bilingual Education Programs. Palo Alto, CA, 1975.

American Institutes for Research. Evaluation of the Impact of ESEA Title VII Spanish/English Bilingual Education Program, Vol. I: Study Design and Interim Findings. Palo Alto, CA, 1977.

American Institutes for Research. Evaluation of the Impact of ESEA Title VII Spanish/English Bilingual Education Program, Vol. II: Project Descriptions. Palo Alto, CA, 1977.

American Institutes for Research. Evaluation of the Impact of ESEA Title VII Spanish/English Bilingual Education Program: Overview of Study and Findings. Palo Alto, CA, 1978.

American Institutes for Research. Evaluation of the Impact of ESEA Title VII Spanish/English Bilingual Education Program, Vol. III: Year Two Impact Data, Educational Process, and In-Depth Analyses. Palo Alto, CA, 1977.

Anastasi, Anne. Psychological Testing, 4th ed. New York: MacMillan, 1976.

Arizona Occupational Research Coordination Unit, Research and Analysis of Competencies Needed by the Bilingual Teacher Aide. Phoenix: Author (1976).

BABEL (Bay Area Bilingual Education League). Evaluation Instruments for Bilingual Education: An Annotated Bibliography. Bilingual Testing and Assessment: Proceedings of Babel Workshop and Preliminary Findings -- Multilingual Assessment Program. Berkeley, CA: BABEL, Dissemination Center for Bilingual-Bicultural Education, 1975.

Berliner, D.C. and Tikunoff, W.J. "The California Beginning Teacher Evaluation Studies: Overview of the Ethnographic Study", Journal of Teacher Education 27, 1, p.24-34 (1976).

Bernal, E.M., Jr. "Assessment Procedures for Chicano Children: The Sad State of the Art", Aztlan, 8, 1977. Bruck, M., Shultz, J., Rodriguez-Brown, An Ethnographic Analysis of the Language Use Patterns of Bilingually Schooled Children. Paper.

Bilingual Education Applied Research Unit. Hunter College. Sources of Information on Assessment for Bilingual Programs. New York: Hunter College, 1975.

Blanco, George. Survey of Measures of Linguistic Proficiency. Austin, Texas: Texas Education Agency/University of Texas at Austin, 1977.

Bloomsbury West, "Learning and Teaching Research in Bilingual Education," Paper for N.I.E., San Francisco: Bloomsbury West, 1977a.

Borich, G.D., and Madden, S.R. Evaluating classroom instruction: A sourcebook of instruments. Reading, MA: Addison-Wesley Publishing Co., Inc. 1977.

Borich, G.D., and Malitz, D. convergent and discriminant validation of three classroom observation systems: A proposed model. Austin, TX: University of Texas at Austin, Research and Development Center for Teacher Education, 1975.

Borich, G.D., Malitz, D., Kugle, C.D., and Pascone, M. Convergent and discriminant validity of five classroom observtion systems: Testing the model. Paper presented at the annual meeting of the American Educational Research Associatin, New York, 1977.

Briere, Eugene J. Testing Student Performance in Phonology. Paper for the NAFSA Conference at Houston, April 1967.

Bruck, M., Shultz, J., Rodriguez-Brown, An Ethnographic Analysis of the Language Use Patterns of Bilingually Schooled Children. Paper.

Buros, Oscar K. (ed.) Mental Measurements Yearbook, 8th Ed. Highland Park, NJ: Grypton Press, 1978.

Burt, M. and H. Dulay, "Some Guidelines for the Assessment of Oral Language Proficiency and Dominance", TESOL Quarterly, 12, 2, p. 177, 1978.

Byers, P. and H. Byers, "Non-Verbal Communication in the Education of children" Functions of Language in the Classroom. Edited by C. Cazden, V. John, and D. Hymes. New York: Teachers College Press (1972).

Calkins, D.S., Borich, G.D., Pascone, M., and Kugle, C.L. Generalizability of teacher behaviors across classroom observation systems. Paper presented at the annual meeting of the American Educational Research Association, New York, 1977.

Campbell, D.T "Degrees of Freedom" and the Case Study", Qualitative and Quantitative Methods in Evaluation Research, Beverly Hills: Sage (1979).

Campbell, D.T., and Fiske, D.W. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 1959, 56, 81 -0105.

Campbell, Donald T. "Degrees of Freedom and the Case Study," Sage Research Progress Series in Evaluation, Vol. 1, 1979.

Campeau, P.L., A.O.H. Roberts, J.E. Bowers. M. Austin, S.J. Roberts, The Identification and Description of Exemplary Bilingual Education Programs, Palo Alto: American Institutes for Research (1975).

Cardenas, Jose. "AIR Evaluation of Bilingual Education," Intercultural Development Research Association (IDRA). Newsletter, June, 1977.

Carroll, John B. "Fundamental Considerations in Testing for English Language Proficiency of Foreign Students," in Testing, Center for Applied Linguistics, 1961.

Center for Applied Linguistics (CAL). "Response to AIR Study Evaluation of the Impact of ESEA Title VII Spanish/English Bilingual Education Programs," Arlington, VA, April 18, 1977.

Center for Applied Linguistics (CAL). "Title VII: A Failure? Some Experts Disagree." The Linguistic Reporter, No. 1, Vol. 21, September, 1978.

Center for Applied Linguistics, Challenge to USOE final evaluation of the impact of ESEA Title VII Spanish/English bilingual educational programs. Duplicated. Arlington, VA. Center for Applied Linguistics, 1978.

Center for Bilingual Education, Northwest Regional Educational Laboratory. Assessment Instruments in Bilingual Education, A Descriptive Catalog of 342 Oral and Written Tests. Los Angeles, California: National Dissemination and Assessment Center, California State University, 1978.

Cervantes, Robert A. "An Exemplary Consafic Chinatropic Assessment: The AIR Report," paper presented at the Association of Mexican American Educators, San Francisco, CA, October 26-28, 1978.

Cervenka, Edward J. Inventory of Socialization for Bilingual Children. Austin, Texas: Child Development Evaluation and Research Center, The University of Texas at Austin, 1968.

Cohen, J. Statistical Power Analysis for the Behavioral Sciences. New York: Academic Press (1977)

Cohen, S.A. and J.S. Hyman. "How Come So Many Hypotheses In Educational Research Are Supported? (A Modest Proposal)." Educational Researcher, 8, 11, pp. 12-16, 1979.

Cook, Thomas D. and Charles S. Reichardt, eds. "Qualitative and Quantitative Methods in Evaluation Research," in Sage Research Progress Series in Evaluation, Vol. 1, 1979.

Cooley, W.W. and Leinhardt, G., Design for the individualized instruction study: A study of the effectiveness of individualized instruction in the teaching of reading and mathematics in compensatory education programs. Pittsburgh: University of Pittsburgh, Learning Research & Development Center, 1975. (b).

Cooley, William W., and Lohnes, P.R., Evaluation Research in Education. New York: John Wiley & Sons, Inc., 1976.

Cooper, E. "Test Selection in Bilingual Education Evaluation," paper presented at the American Educational Research Association Convention, 1978.

Cronbach, L.J. and Furby, L. "How We Should Measure 'Change'--Or Should We?" Psychological Bulletin, 74, 1, pp. 68-80. 1970.

Cronbach L.J. et an. The Dependability of Behaviorally of Behavioral Measurements: Theory of Generalizability for Scores and Profiles, John Wiley and Sons, Inc., USA, 1972.

Cronbach, L. J., Rajaratnam, M., and Gleser, G., Theory of generalizability: A liberalization of reliability theory. British Journal of Statistical Psychology, 1963, 16, 137 - 163.

Cummins, J. "Linguistic Interdependence and Educational Development of Bilingual Children." Review of Educational Research, 49, 2, pp. 222-251, 1979.

Cummins, James. "The Cognitive Development of Bilingual Children: A Review of Recent Research," 1977.

Danoff, M. et al. Evaluation of the Impact of ESEA Title VII Spanish/ English Bilingual Education Program, Volume I: Study Design and Interim Findings, AIR Report 48300-2-77-FRIII (Palo Alto, Calif.: AIR, February 1977), ERIC report ED 138-090.

Danoff, M. et al. Evaluation of the Impact of ESEA Title VII Spanish/ English Bilingual Education Program, Volume II: Project Descriptions, AIR Report 48300-2-77 FRIV (Palo Alto, Calif.: AIR, February 1977), ERIC Report ED 138-091

Danoff, M.N. et al., Evaluation of the Impact of ESEA Title VII Spanish/ English Bilingual Education Program, Overview of Study and Findings, AIR Report 48300-3-78-FRVI (Palo Alto, Calif.: AIR, March 1978), ERIC Report ED 154-634.

Danoff, M.N., G.J. Coles, D. McLaughlin, and D. Reynolds. Evaluation of the Impact of ESEA Title VII Spanish/ English Bilingual Education Program, Volume III: Year Two Impact Data. Educational Process, and In-Depth Analyses, AIR Report 68300-1-78-FRV (Palo Alto, Calif.: AIR, January 1978), ERIC Report ED 154-634.

DeGeorge, George P., Selective Classified Test List for Spanish-Speaking Bilingual Students, Cambridge, Maine: Lesley College National Assessment and Dissemination Center for Bilingual-Bicultural Education, 1976.

Deutsch, Martin, Joshua A. Fishman, Leonard S. Kogan, Robert D. North, and Martin Whitman. "Guidelines for Testing Minority Group Children," in The Journal of Social Issues, 1964, Vol. 20.

Dissemination Center for Bilingual Bicultural Education, Evaluation Instruments for Bilingual Education: An Annoted Bibliography, Austin, Texas, 1975.

Dissemination Center for Bilingual Bicultural Education. Evaluation Instruments for Bilingual Education: A Review of Tests in Use in Title VII Bilingual Education Projects. Austin, Texas, n.d.

Division of Evaluation, New York State Education Department, Bilingual Education Evaluation (Spanish-English), Resources for the Selection and Application of Evaluation Instruments to Measure Results of Bilingual and English As a Second Language Programs. Albany, NY: The University of the State of New York, 1975.

Dulay, H. and Burt, M. "Bilingual Education: A close look at its effects", FOCUS, Arlington, VA: National Clearinghouse for Bilingual Education, n.d.

Dulay, H., and Burt, M., Why Bilingual Education? A Summary of Research Findings, Poster, San Francisco, Boomsbury West, 1977b.

Dulay, Heidi and Marina Burt. "The Efficacy of Bilingual Education" in Educational Evaluation and Policy Analysis Sept.-Oct. 1979, No. 5.

Dulay, Heidi, Marina Burt and Loraine Zappert. Why Bilingual Education: A Summary of Research Findings. Berkeley, CA: BABEL/Lau Center, 1976.

Ebel, R. "Estimation of the reliability of ratings." Psychometrika, 1951, 16 (4), 407 - 424.

Ehrlich, Allan and Roselin Ehrlich (editors), Tests in Spanish and Other Language, and Non-Verbal Tests for Bilingual: An Annotated B.E.A.R.U. Bibliography, New York: Hunter College, Project BEST, Bilingual Education Applied Linguistics Unit, 1974.

Ehrlich, Allan and Roselin Ehrlich (eds.) Tests in Spanish and Other Languages, English as a Second Language, and NonVerbal Tests for Bilinguals: An Annotated B.E.A.R.U. Bibliography. New York: Hunter College, Project BEST, Bilingual Education Applied Linguistics Unit, 1974.

Erlich, O., and Shavelson, R. The application of generalizability theory to the study of teaching. Beginning Teacher Evaluation Study, Technical Report No. 76-9-1. San Francisco, CA: Far West Laboratory for Educational Research and Development, 1976.

Fishman, J., Language Loyalty in the United States, The Hague: Mouton, 1966

Fishman, J., "The status and prospects of bilingualism in the United States", Modern Language Journal, 49 (3), p. 143-55, 1975.

Fishman, J. A. "The Significance of the Ethnic Community Mother Tongue School: Introduction to a Study," NABE Journal III, 3, Spring 1979.

Flanders, N.A. Interaction analysis in the classroom: A manual for observers. In A. Simon and E.G. Boyer (Eds.), Mirrors for Behavior Philadelphia, PA: Research for Better Schools, 1967.

Frick, T. and Semmel, M.I. Observer agreemunt and reliabilities of classroom measures. Review of Educational Research, 1978, 48 (1), 157-184.

General Accounting Office. Bilingual Education: An Unmet Need. Report to the Congress by the Comptroller General of the United States, Washington, D.C., 1976.

Gezi, Kal. "Bilingual-Bicultural Education: A Review of Relevant Research," in California Journal of Educational Research, 1975, Vol. XXV, No. 5, pp. 223-239.

Glaser, B., and Strauss, A.L., The Discoveries of Grounded Theory, Chicago: Oldine (1967).

Glass, G.V. "Integrating Findings: The Meta-Analysis of Research." Review of Research in Education, 5, pp. 351-379, 1978.

Glass, G.V. "Primary, Secondary, and Meta-Analysis of Research," Educational Researcher, pp. 3-8, November 1976.

Glass, G. V. "Primary, Secondary and Meta-analysis of Research," Review of Research in Education, No. 5, 1978.

Glass, G.V. and M.L. Smith. "Meta-Analysis of Research in Class Size and Achievement." Educational Evaluation and Policy Analysis, 1, 1, pp. 2-16, 1979.

Glyn-Lewis, E., Report on the Seminar: Bilingualism in Education, International Seminar, Aberystwyth, Wales, 1960; London: Her Majesty's Stationery Office, 1965.

Gray, Tracy C. "Challenge to USOE Final Evaluation of the Impact of ESEA Title VII Spanish/English Bilingual Education Programs." Arlington, Virginia: Center for Applied Linguistics, 1977.

Gray, Tracy C., Response to AIR study. Duplicated. Arlington, VA, Center for Applied Linguistics, 1977.

Gutierrez, Medardo, John Rosenbach and Gustavo Gonzalez. Bilingual Assessment Test Development and Review: A Manual for Teacher Use. Albany, NY: The State Education Department, Office of Research, Planning and Evaluation, Office of Evaluation, 1975.

Haro, C.M. "Truant and Low-Achieving Chicano Student Perceptions in the High School System", Aztlan, 8,01977.

Holtzman, Paul D. "English Language Proficiency Testing and the Individual." Workpaper on testing, ATESL Seminar, Houston, Texas, April 1967.

Horn, Thomas D. and Carmela Hernandez. "Standardized Testing and the Spanish-Speaking Minorities: Looking Beyond the Correlation."

Hymes, Dell. "Introduction: Toward Ethnographies of Communication," in American Anthropologist, Special Publication, ed. by John J. Gumperz and Dell Hymes. Vol. 66, No. 6, Part 2, Dec., 1964.

Jackson, G., and Cora C. "The Inequality of Educational Opportunity in the Southwest: An Observational Study of Ethnically Mixed Classrooms." American Educational Research Journal, 11, pp. 219-229, 1974.

Kane, M.T. and R.L. Brennan, "The Generalizability of Class Means." Review of Educational Research, 47, 1977, pp 267 - 293.

Kane, M.T., G.M. Gilmore and T.J. Crooks. "Student Evaluations of Teaching: The Generalizability of Class Means," Journal of Educational Measurement, 13, 1976, pp. 171-183.

Kennedy, M.M. "Generalizing for Single Case Studies," Evaluation Quarterly, 3, 4, pp. 661-678, 1979.

Kenny, D.A. "Cross-lagged and Synchronous Common Factors in Panel Data," Chapter 8 of Structional Equation Models in the Social Sciences, Goldberger and O.D. Duncal, eds. New York: Seminar Press, 1973.

Laosa, L.M. Inequality in the Classroom: Observational Research on Teacher-Student Interactions, Aztlan, 8, 1977.

Leinhardt, G. "Applying a Classroom Process Model to Instructional Evaluation." Curriculum Inquiry, 8, 1978b, pp 155-176.

Leinhardt, G. "Coming Out of the Laboratory Closet," Social Psychology of Education: Theory and Research. D. Bart-Tal and L. Saxe (eds).

Leinhardt, G., Measurement of Treatment, Draft Paper,Pittsburgh: Author (1978).

Leinhardt G. Program evaluation: An empirical study of individualized instruction. American Educational Research Journal, 1977, 14(3, 277 - 293.

Leopold, Werner F. Speech Development of a Bilingual Child: A Linguist's Record. (Northwestern University Studies, Humanity Series, Nov. 6, 11, 18, 19.) Evanston, Ill.: Northwestern University Press, 4 vols., 1939-1949.

Light, R. L. "Capitalization on Variation: How Conflicting Research Findings Can Be Helpful for Policy," Educational Researcher, Vol. 8, 9, 1979.

Light, R.L. "Issues in Teacher Preparation for Cross-Cultural Education", paper presented at the International Conference on Bilingual-Bicultural Education, New York, 1974.

Llanes, J.R. "Human Enculturation Within Group Culture-Clusters" Eastern Michigan University: Ypsilanti (1979).

Llanes, J.R., R. Cruz, and A. Vega, "A Classroom-Based Approach to Research in Bilingual Education Processes", Journal of Chicano Research, p.107, 1976.

Locks, Nancy A., Barbara A. Pletcher, and Dorothy F. Reynolds. Language Assessment Instruments for Limited English-Speaking Students: A Needs Analysis, Washington, D.C.: National Institute of Education, 1978.

Marliave, R., Fisher, C., Filby, ., and Dishaw, M. The development of instrumentation for a field study of teaching. Beginning Teacher Evaluation Study, Technical Report 1-5. San Francisco, CA: Far West Laboratory for Educational Research and Development, 1977.

McDavid, Raven I., Jr. and William M. Austin. "Communication Barriers to the Culturally Deprived," USOE Cooperative Research Project 2107, n.d.

McGaw, B., Wardrop, J.L., and Bunda, M.A. Classroom observation schemes: where are the errors? American Educational Research Journal, 1972, 9(1), 13-27.

Medley, D.M. Teacher competence and teacher effectiveness: A review of process-product research. Washington, DC: American Association of Colleges for Teacher Education, 1977.

Medley, D.M., and Mitzel, H.E. Assessing the learning environment in the classroom: A manual for OSCAR 5V users. Unpublished manuscript, University of Virginia, n.d.

Medley, D.M., and Mitzel, H.E. Application of analysis of variance to the estimation of the reliability of observations of teachers' classroom behavior. Journal of Experimental Education, 1958, 27, 23-35.(a).

Medley, D.M., and Mitzel, H.E. A technique for measuring classroom behavior. Journal of Educational Psychology, 1958, 49 (2), 86-92.(b).

Medley, D.M., and Mitzel, H.E. Measuring classroom behavior by systematic observation. In N.L. Gage (Ed.) Handbook of Research and Teaching. Chicago, IL: Rand-McNally, 1963, 247-328.

Mehan, H. "Pasticipation Structures", paper prepared for Far West Laboratory for Educational Research and Development, January 1979.

Mehan, H. "Structuring School Structures", Harvard Educational Review, 48, 1, pp. 32-64, February 1978.

Mehan, H. "The Structure of Classroom Events and Their Consequences for Students' Performance", in Children In and Out of School, Philadelphia: University of Pennsylvania, (1977).

Mosteller, F. and Tukey, J.W. Data Analysis and Regression, Reading, MA: Addison Wesley (1977).

The National Advisory Council on Bilingual Education. Annual Report. Washington, D.C.: Office of Bilingual Education, 1975, 1976 and 1977.

National Dissemination and Assessment Center. Assessment Instruments in Bilingual Education: A Descriptive Catalog of 342 Oral and Written Tests. Los Angeles, CA: California State University, 1978.

Nickel, K.N. "Experimentation, Extrapolation, Exaggeration: Thy Name is Research", Phi Delta Kappa, p.260-61, December 1979.

North, Robert D. and Edward J. Cervenka. Evaluation Report for the Project: A Program for Pupils in Non-Public Schools Learning English as a Second Language. New York: The Psychological Corp., 1971.

North, Robert D. and Edward J. Cervenka. Final Evaluation Report for the English as a Second Language Program, Boston Public Schools. New York: The Psychological Corp., 1972.

Nunnally, J. C., Psychometric Theory. New York: McGraw Hill, 1967.

Office for Civil Rights, Directory of Elementary and Secondary School Districts, and Schools in Selected School Districts School Year 1976-1977, Washington: DHEW/OCR (ND).

O'Malley, Michael J. "Review of Interim Report: Evaluation of the Impact of ESEA Title VII Spanish/English Bilingual Education Programs." Vols. I and II. (Memorandum to John Molina), NIE, April 22, 1977.

O'Malley, Michael J. "Testimony for the Subcommittee on Elementary and Secondary and Vocational Education of the Committee on Education and Labor, U.S. House of Representatives," June 7, 1977.

Peal, Elizabeth and Wallace E. Lambent. "The Relation of Bilingualism to Intelligence," Psychological Monographs: General and Applied, Vol. LXXVI, No. 27, 1962.

Pletcher, Barbara P., Nancy A. Locks, Dorothy F. Reynolds, and Bonnie G. Sisson. A Guide to Assessment Instruments for Limited English Speaking Students. New York: Santillana Publishing Co., Inc., 1978.

Ramirez, M., J. Ramirez, T. Leaffer, and M. Silva, Bilingual/Bicultural Early Educationl Development Research Workshop Proceedings, Washington: DHEW, Office of Child Development.

Reichardt, C.S., and T.D. Cook, "Beyond Qualitative versus Quantitative Methods," Qualitative and Quantitative Methods in Evaluation Research, Beverly Hills: Sage (1979).

Rodriguez-Brown, F.V. et.al., "Language Interaction in a Bilingual Classroom: An Observational Study", paper presented at 1976 meeting of A.E.R.A., San Francisco, 1976.

Rodriguez-Brown, F.V. The Effect of Language Used for Early Reading Instruction: A Bilingual Perspective, Paper presented at the 1979 meeting of A.E.R.A., San Francisco, 1979.

Rodriguez-Brown, F.V., M. Bruck, C.S. Walcer, and J. Shultz, Language Interaction in a Bilingual Classroom: An Observational Study, Paper presented at the 1976 Meeting of A.E.R.A., San Francisco, 1976.

Rosenshine, B., and Furst. N. The use of direct observation to study teaching. In R.M.W. Travers (Ed.), Second Handbook of Research on Teaching. Chicago, IL: Rand-McNally, 1973.

Rowley, G. The reliability of observational measures. American Educational Research Journal, 1976, 13(1), 51-59.

Rowley G. The relationship of reliability in classroom research to the amount of observation: An extension of the Spearman-Brown formula. Journal of Educational Measurement, 1978, 15(3), 165-180.

Rubovits, P., and M. Maehe "Pygmalian Black and White". Journal of Personality and Social Psychology 25, pp.210-218, 1973.

Salamanca, A.J., "Bilingual/Cross-Cultural Teacher Shortage in California. Position Paper", Paper presented at Bilingual/Cross-cultural Teacher Preparation Workshop, San Diego, 1974.

Sanchez, R., H. Romo, Santos Rivera,I, and B. Williams, Issues in Language Proficiency Assessment, San Diego: Institute for Cultural Pluralism, San Diego State University ,1978).

Savard, Jean-Guy. Bibliographie Analytique de Tests de Langue/Analytical Bibliography of Language Tests. Quebec: International Center for Research on Bilingualism Les Presses de L'Universite Laval, 1972.

Scott, W.A. Reliability of content analysis: The case of nominal scale coding. Public Opinion Quarterly, 1955, 19, 321-325.

Sherry, R.W. "Toward a Cross-Disciplinary View", NABE Journal, III, 1, Fall, 1978.

Silverman, Robert J., Joslyn K. Noa, and Randall H. Russell. Oral Language Tests for Bilingual Students: An Evaluation of Language Dominance and Proficiency Instruments, Portland, Oregon: Center for Bilingual Education, Northwest Regional Educational Laboratory, 1976.

Simon, A., and Boyer, E.G. (Eds.) Mirrors for Behavior: An Anthology of Observation Instruments. Philadelphia, PA: Research for Better Schools, 1970.

Skutnabb-Kangas, Tove, and Pertti Toukomaa. Teaching Migrant Children's Mother Tongue and Learning the Language of the Host Country in the Context of the Socio-Cultural Situation of the Migrant Family. Helsinki: Finnish National Commission for UNESCO, 1976.

Stufflebeam, D. L. et. al. Educational Evaluation and Decision Making. Itasca, Ill., Peacock, 1971.

Tamera, E. "Bilingual-Bicultural Education Today: State of the Art," in General Information about Bilingual Education, Austin: Bilingual Center, ESC, Region XII, n.d.

Taylor, Marie E. An Overview of Research on Bilingualism. Sacramento, Cal., California State Department of Education, 1970.

Thorndike, R. L. and Hagen, E. Measurement and Evaluation in Psychology and Education. New York: John Wiley, 1969.

Tikunoff, W.J., Berliner, D.C., and Rist, R.C. Abstract from Special Study A: An ethnographic study of the forty classrooms of the Beginning Teacher Evaluation Study known sample. Beginning Teacher Evaluation Study, Technical Report No. 75-10-5-(A). San Francisco, CA: Far West Laboratory for Educational Research and Development, 1975.

Tikunoff, W. J., C. C. Berliner, and R. C. Rist. Sample of Fifth Grade Classroom Protocols, Far West Laboratory for Research and Development, San Francisco, 1975.

Toukomaa, Pertti, and Tove Skutnabb-Kangas. The Intensive Teaching of the Mother Tongue to Migrant Children of Pre-School Age. Research Reports 26. Tampere, Finland: Department of Sociology and Social Psychology, University of Tampere, 1977.

GPO 857-094